Analyse des leviers : effets de colinéarité et hiérarchisation des impacts dans les études de marché et sociales

Henri Wallard

THÈSE

présentée par **Henri WALLARD**

soutenue le : 18 décembre 2015

pour obtenir le grade de :

Docteur du Conservatoire National des Arts et Métiers

Discipline/ Spécialité : Informatique

Analyse des leviers

Effets de colinéarité et hiérarchisation des impacts dans les études de marché et sociales.

THÈSE dirigée par :

M. Gilbert SAPORTA — Professeur émérite, CNAM, Paris.

RAPPORTEURS :

M. Julien JACQUES — Professeur, Université de Lyon-Lumière, Lyon.

M. Jean-Michel POGGI — Professeur, Université Paris Descartes et Laboratoire de Mathématiques d'Orsay (LMO), Université Paris-Sud, Orsay.

JURY présidé par :

M. Michel DELECROIX — Professeur, Université Pierre et Marie Curie, Paris.

JURY :

M. Aurélien LATOUCHE — Professeur, CNAM, Paris.

M. Mohamed NADIF — Professeur, Université Paris Descartes, Paris.

A mes parents, Edmée et Paul.

Remerciements

Mes remerciements vont tout d'abord d'abord au Professeur Gilbert Saporta qui a bien voulu diriger cette recherche, et m'a apporté au long de ces trois années des conseils et orientations extrêmement précieux, faisant de ce travail à la fois une chance et une aventure passionnante.

Je remercie les Professeurs Julien Jacques et Jean-Michel Poggi qui m'ont fait l'honneur d'accepter d'être les rapporteurs. Le temps qu'ils ont bien voulu accorder à la relecture de ce manuscrit a permis de rectifier la forme, d'améliorer la clarté de plusieurs parties du texte et de compléter utilement certains points. Je remercie les Professeurs Michel Delecroix, Aurélien Latouche et Mohamed Nadif d'avoir accepté de faire partie du jury.

Je tiens également à remercier Ndeye Niang, Michel Béra, Pierre-Louis Gonzalez, Giorgio Russolillo et toute l'équipe du Cédric pour leur accueil chaleureux et leur grande gentillesse. Philippe Périé m'a apporté son expertise en statistique et en informatique qui m'a été très utile pour organiser la programmation avec R. Les discussions multiples et enrichissantes avec Antoine Moreau ont participé à ma motivation et Hervé Mignot a toujours été disponible pour des échanges stimulants. Je les en remercie.

Un dernier mot pour Jean-Paul Aimetti qui m'a encouragé avec bienveillance à me lancer dans ce projet.

Résumé

Les praticiens des études de marché et sociales cherchent à hiérarchiser et quantifier les impacts de leviers d'action sur un objectif recherché. Ceci peut être appliqué en vue d'identifier des priorités d'action parmi différents éléments d'un service ou d'un produit afin d'améliorer la satisfaction des clients. La régression linéaire multiple est l'une des méthodes les plus utilisées pour établir cette hiérarchisation et aussi pour simuler les impacts des différents prédicteurs. Cependant les études par enquêtes comportent souvent des questions sur des sujets très voisins conduisant à une proximité des réponses obtenues et donc à une colinéarité entre les prédicteurs. Cette colinéarité entraîne la possibilité d'obtenir des coefficients négatifs contraires à l'intuition des experts et aussi de grands intervalles de confiance des coefficients estimés qui traduisent une instabilité des résultats et une non-reproductibilité entre les enquêtes.

La quantification de l'importance des prédicteurs a été étudiée notamment depuis 1960 pour tenter de trouver des réponses appropriées à ces difficultés et reste en un sujet de recherche actif (Grömping (2015)).

Nous avons analysé trois approches : les méthodes d'estimations de valeurs d'importance à partir du modèle linéaire, la quantification de l'importance avec les forêts aléatoires et l'utilisation des réseaux bayésiens.

Une revue des méthodes d'allocation de la variance expliquée référencées dans la littérature scientifique et l'industrie des études de marché a été réalisée et douze méthodes fondées sur le modèle linéaire ont été considérées dans cette recherche parmi lesquelles neuf consistent en une décomposition de la variance expliquée. Les liens entre certaines méthodes sont présentés et une démonstration de l'égalité entre la méthode *lmg-Shapley* et la méthode de Johnson dans le cas avec deux prédicteurs est proposée via une formalisation trigonométrique.

Plusieurs résultats nouveaux ont également été obtenus, notamment que la méthode de Fabbris (1980) n'est pas identique aux méthodes de Genizi (1993) et Johnson (2000) et que la méthode des *CAR scores*, parfois aussi nommée méthode de Gibson (1962), ne conduit pas nécessairement à des scores identiques pour deux prédicteurs lorsque leur corrélation tend vers 1. Enfin une méthode nouvelle, baptisée *weifila* (*weighted first last*) beaucoup moins gourmande que *lmg-Shapley* en termes de calcul a été proposée qui conduit à des résultats extrêmement proches et même identiques dans le cas de deux prédicteurs. La méthode *weifila* a fait l'objet d'une publication en 2015 figurant en annexe.

En ce qui concerne les forêts aléatoires et l'importance des variables par permutations, les résultats obtenus conduisent à considérer que les mérites de ces approches sont sous-estimés. Il a ainsi été observé que le choix de faibles valeurs de mtry dans l'application des forêts aléatoires avec la méthode RF-CART permet de calculer des répartitions d'importance très proches des résultats obtenus avec *lmg-Shapley* ou la méthode de Johnson mais avec l'avantage d'une excellente prise en compte des non-linéarités. Ce résultat a été établi à la fois en reprenant avec une approche différente des données déjà utilisées dans la littérature (Grömping (2009 et 2015)) mais aussi en employant des données simulées. A ce titre les forêts aléatoires, malgré leur caractère non-paramétrique et leur efficacité en termes prédictifs et de hiérarchisation, restent quasiment inexistantes dans les études de marché et mériteraient une utilisation plus large.

A la différence de publications antérieures, notre recherche conduit donc à ne pas proposer *lmg-Shapley* ou *pmvd* comme méthodes de références, mais soit *weifila* si la rapidité et la simplicité sont recherchées, soit les forêts aléatoires afin de bien prendre en compte les non-linéarités et interactions.

En ce qui concerne l'utilisation des réseaux bayésiens, qui ont connu une forte popularité ces dernières années, la multiplicité des solutions possibles et le recours fréquent à des restrictions et choix d'expert dans l'élaboration des structures des graphes militent pour une prudence dans la communication des mérites de ces méthodes notamment quant à la découverte possible de relations causales. Ils apportent cependant un moyen intéressant d'explorer des modèles possibles et de réaliser des simulations.

Enfin, malgré les mérites des méthodes d'allocation de variance, des forêts aléatoires ou des réseaux bayésiens, il reste important de ne pas négliger l'intérêt des approches structurelles et l'apport des modèles conceptuels.

Mots clés : *régression, décomposition de la variance, importance, valeur de Shapley, forêts aléatoires, réseaux bayésiens.*

Abstract

Market research practitioners want to identify a ranking of drivers of action or quantify of their respective impact on a desired outcome. This may consist in identifying priorities of action in the service or product mix to increase customer satisfaction. Linear regression is one of the most popular method to rank drivers and model their respective impact. But market research surveys often include questions that relate to similar topics and as a consequence the answers provided by respondents translate into correlated predictors. This multicollinearity creates circumstances where regression coefficients can be difficult to use directly because they may have negative values that are counterintuitive for the experts or they can be instable across samples, and the results of the survey cannot be reproduced.

Variable importance has been investigated in particular since 1960 to try to find adequate methods and remains an active field of research (Grömping (2015)).

We have analyzed three main approaches: variable importance derived from linear model results, permutation importance with random forests, and Bayesian networks.

A review of the methods to allocate explained variance in the scientific and market research literature has been conducted and twelve methods based on linear model have been considered, among which nine consist in a full decomposition of the explained variance. Relationships between various methods are presented and a trigonometric demonstration of the full equality between Johnson and *lmg-Shapley* decomposition in the case of two predictors has been proposed.

Several new results have been identified, among which the fact that the decomposition proposed by Fabbris (1980) is not identical to the method proposed by Genizi (1993) and Johnson, and also that *CAR scores*, also called Gibson method, do not necessarily tend to equate when the correlation between two predictors tend towards 1. A new method to decompose variance has been introduced called *weifila* for *weighted first last,* and this method is much less computer intensive than *lmg-Shapley*, is equal to Johnson and *lmg-Shapley* in the case of two predictors and gives very similar results when there are more than two predictors. This method has been presented in a publication in 2015 that is shown in the appendix.

Regarding variable importance estimation using random forests, the benefits of this method appear to be underestimated. We have observed that if low values of mtry are chosen for Random Forest with RF-CART, the importance values are allocated in very similar proportions than using *lmg-Shapley* or Johnson but with the additional benefit to efficiently take into account non linearity. This result has been validated using in a different way data used in prior publications (Grömping (2009) and (2015)) but also using simulated data. The interest of this approach in marketing research applications should be reinforced and the usage of random forests should be expanded.

This research proposes different conclusions than prior work as to the preferred methods to use for variance decomposition leading to not recommend *lmg-Shapley* or *pmvd* and to propose instead *weifila* for its simplicity or random forests because of the possibility to handle interactions and non-linearity.

Regarding Bayesian networks, which have become very popular these last few years, the multiplicity of methods and the necessity to make assumptions and expert driven decisions in the discovery of arcs should lead to very cautious regarding the publicised capacity of Bayesian networks to justify causal links, even if they represent a useful tool to explore data structures and to implement simulations.

Whatever are the merits of variance decomposition random forests and Bayesian networks this should not lead to disregard the benefits of pre-agreed models based on theory.

Key words : *Regression, Variable Importance, Variance Decomposition, Shapley Value, Random Forests, Bayesian Networks.*

Table des matières

Introduction

Les études de marché et les études sociales font appel à des modèles statistiques afin de déterminer quelles actions doivent être mises en œuvre en priorité pour obtenir un résultat donné, comme par exemple augmenter la fidélité ou la satisfaction de clients et aussi pour simuler un impact attendu en agissant sur ces leviers. Ainsi les praticiens d'études de marché vont utiliser la régression linéaire pour tenter d'attribuer une importance relative à des leviers d'action. Cependant les corrélations entre prédicteurs créent des difficultés : instabilité dans les coefficients mesurés entre des enquêtes successives comparables, ou encore apparition de coefficients négatifs ce qui peut être contre-intuitif. Cette situation est fréquemment rencontrée dans les études par enquêtes quand des questions voisines (par exemple sur l'attrait d'un produit ou sur une pratique sociale) sont formulées dans les questionnaires présentés aux répondants. Les variables construites à partir des réponses obtenues et utilisées pour la modélisation sont alors corrélées.

Notre recherche porte principalement sur les différentes méthodes utilisées afin de quantifier l'importance des prédicteurs, soit par allocation aux prédicteurs de parts de la variance expliquée, soit par l'utilisation des forêts aléatoires. Elle donnera également quelques perspectives sur l'utilisation des réseaux bayésiens.
Après une analyse de certaines recommandations et de l'utilisation de packages statistiques publiés depuis une dizaine d'années cette recherche présente plusieurs résultats théoriques et formule des recommandations spécifiques pour l'estimation pratique de l'importance des prédicteurs.

Notations et rappels

Les matrices sont désignées par des lettre majuscules à caractère gras (exemple : $\mathbf{Z}$), les vecteurs par des lettres minuscules à caractère gras (exemple : $\mathbf{y}$), les éléments des vecteurs et matrices par des lettres minuscules en italique avec les indices appropriés si nécessaire (exemple $z_{i,j}$ est un élément de $\mathbf{Z}$).

$\mathbf{Z'}$ désigne la matrice transposée de la matrice $\mathbf{Z}$.

La décomposition en valeurs singulières de la matrice $\mathbf{X}$ sera notée $\mathbf{P\Delta Q'}$.

Le produit de Hadamard de deux matrice (k,l) $\mathbf{M}$ et $\mathbf{N}$ sera noté $\mathbf{M.N}$. Rappelons que le produit matriciel de Hadamard est une opération qui pour deux matrices de mêmes dimensions associe une autre matrice, de même dimension où chaque élément de la matrice produit est le produit terme à terme des éléments des deux matrices.

Ainsi en notant $\mathbf{P = M.N}$, $p_{i,j} = m_{i,j} n_{i,j}$.

La régression linéaire appliquée aux résultats d'enquêtes est présentée avec les notations suivantes.

Considérons que nous avons observé sur n individus $p+1$ variables représentées par des vecteurs de $\mathbf{R}^n$ $\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_p}$,

$\mathbf{y}$ est désignée comme la variable à prédire et les $\mathbf{x_j}$ sont les prédicteurs.

$r(\mathbf{y}, \mathbf{x_i})$ désigne le coefficient empirique de corrélation entre les deux variables $\mathbf{y}$ et $\mathbf{x_i}$

Soient :

- $\mathbf{X}$ la matrice à n lignes dont les colonnes sont $\mathbf{1}, \mathbf{x_1}, ..., \mathbf{x_p}$
- $\mathbf{b}$ la solution des moindres carrés $\mathbf{b = (X'X)^{-1} X'y}$
- $\mathbf{y^* = Xb}$

R désigne le coefficient de corrélation entre $\mathbf{y}^*$ et $\mathbf{y}$.

$V(\mathbf{y}^*)$ désigne la variance empirique de $\mathbf{y}^*$.

Le coefficient de régression standardisé pour la variable $\mathbf{x_j}$ est défini comme $\beta_j = b_j \frac{s_j}{s_y}$, s_j étant l'écart-type des observations de la variable $\mathbf{x_i}$.

Nous ferons aussi appel dans le corps du texte de la thèse à la variance expliquée par la regression linéaire en utilisant une partie seulement des variables.

Soit $P = \{1, ..., p\}$ et $K \subset P$, $R^2(K)$ sera le R^2 de la régression linéaire de la variable à expliquer $\mathbf{y}$ sur les variables $\{\mathbf{x_j}\}$ avec $j \in K$. Ainsi par exemple $R^2(1,2)$ est le R^2 obtenu par la régression linéaire de $\mathbf{y}$ sur $\{\mathbf{x_1}, \mathbf{x_2}\}$.

Chapitre 1 : Champ de la recherche

1.1 Les Etudes de Marché

Le champ de cette recherche porte sur la modélisation des leviers dans les études de marché et sociales. Les études de marché représentent une activité économique identifiée, qui est également mise en œuvre par de nombreux acteurs hors de l'industrie des études de marchés au sens strict. Le secteur du « Market Research », tel que défini par les entreprises qui se définissent elles-mêmes comme sociétés d'études de marché, représente un marché mondial estimé à 39 milliards de $ (Rapport Esomar publié en 2013) , et cette estimation inclut pour la première fois une part d' « Advisory Services » à hauteur de 5 milliards de $, reconnaissant de fait l'existence d'activités de Market Research hors des entreprises auto définies comme appartement strictement au secteur du Market Research.

En effet cette activité est aussi pratiquée par de nombreux autres acteurs commerciaux ou académiques : instituts de statistique nationaux, sociétés de conseil, entreprises plus liées à l'informatique (EFM, Entreprise Feedback Management), fondations (eg PEW fondation aux USA) ainsi que de multiples structures de recherche académiques notamment dans le domaine marketing et des études sociales et politiques. Ces activités sont aujourd'hui en mutation en raison du développement de l'essor des applications et services numériques et les frontières entre les différents types d'activités précitées évoluent. Plus récemment le recours au concept de « Big Data » créée de nouvelles opportunités et approches pour fournir de l'information en vue des prises de décision.

Nous parlons ici délibérément d'études de marché et sociales car bien qu'elles soient souvent regroupées dans la définition de ce secteur économique, elles diffèrent souvent en termes de dynamique d'utilisation, de niveau d'exigence méthodologique et de temps consacré aux analyses, dans la mesure où les études sociales sont souvent l'objet d'un investissement en temps et en analyse plus substantiel, et moins automatisé que de nombreuses applications marketing souvent plus tactiques comme par exemple les tests de produits ou les tests publicitaires.

Les études de marché sont définies par l'organisation ESOMAR. ESOMAR est une organisation professionnelle reconnue fondée en 1948 et à laquelle adhèrent de nombreux spécialistes et entreprises du secteur des études de marché autour du monde : *« ESOMAR is the world association of research professionals. Founded in 1948 as the European Society for Opinion and Marketing Research"*. ESOMAR définit les études de marché de la façon suivante :

"Market research, which includes social and opinion research, is the systematic gathering and interpretation of information about individuals or organisations using the statistical and analytical methods and techniques of the applied sciences to gain insight or support decision making. The identity of respondents will not be revealed to the user of the information without explicit consent and no sales approach will be made to them as a direct result of their having provided information". (source: site internet ESOMAR : www.esomar.org)

Les études de marché et sociales sont donc mises en œuvre pour acquérir des informations sur les usages, attitudes et perceptions des consommateurs, clients, citoyens ou plus généralement des représentants d'une population d'intérêt, afin de valider ou d'infirmer des analyses pré existantes , de fournir un état des lieux , mais aussi de permettre d'éclairer des choix possibles d'action , de déterminer des priorités , de quantifier des impacts possibles résultant de la mise en œuvre d'un plan d'action voire de créer un outil de prédiction.

La définition générale d'ESOMAR renvoie à une combinaison d'expertises. Ainsi les termes « gathering and interpretation » requièrent des expertises comme l'échantillonnage ou encore la conception, la rédaction et l'administration du questionnaire auprès des consommateurs ou citoyens interrogés. Cette expertise signifie par exemple être capable de formuler des questions claires et non ambiguës, de savoir ordonner les questions, de ne pas rédiger des questionnaires trop longs de les adapter au mode de recueil par exemple avoir des questions plus courtes si les réponses sont recueillies via une tablette ou un smartphone par comparaison avec le recueil via un PC.

Les éléments théoriques et pratiques de mise en œuvre des études de marché sont décrits dans de nombreux ouvrages, cf. par exemple Giannelloni et Vernette (2012) ou Ardilly (2006). Notre recherche sera centrée sur la partie analytique des études de marché, telle que formulée dans la définition d'ESOMAR à savoir : *« using the statistical and analytical methods and techniques of the applied sciences to gain insight or support decision making ».* Ceci appelle plusieurs commentaires quant au champ de l'étude et aux caractéristiques particulières des variables analysées.

En premier lieu il est important de noter que les variables observées sur des individus sont déterminées par le praticien qui choisit les questions souvent avec l'implication du client de la société d'étude de marché. Cet élément est absolument fondamental comme nous le verrons plus loin dans le choix et l'évaluation des techniques de modélisation. En particulier la colinéarité des variables est le résultat du choix des questions souvent trop nombreuses, ce qui entraine des similarités fortes entre les variables observées, et qui ne permettent en outre pas toujours de couvrir efficacement les dimensions pertinentes du problème étudié. Cet excès de questions a d'autres impacts pénalisants comme la « fatigue » des répondants qui au bout d'un certain temps d'enquête finissent par répondre de façon distraite et automatique aux questions : ce phénomène est appelé « straight lining », en référence au fait que le répondant finit par cocher toujours la même case de réponse sur les échelles d'où une ligne verticale de « croix » par exemple dans le tableau représentant les réponses. Ce phénomène de « straight-lining » est souvent associé une vitesse anormalement élevée des réponses car par exemple dans un questionnaire en ligne le répondant clique mécaniquement rapidement sur la même réponse type (par exemple « moyennement » ou « ne sait pas »).

Dans d'autres cas la lassitude du répondant se manifestera même par un abandon pur et simple de l'enquête en cours de route, c'est le « drop out ». L'accroissement de la durée de réponse à un questionnaire s'accompagne en effet d'un accroissement du taux d'abandon. Il arrive que des questionnaires nécessitent plus de 20 minutes pour être remplis par les répondants (cas de questionnaires en ligne), et au-delà de cette durée le taux d'abandon

augmente de façon nette. Aux inconvénients d'une multi-colinéarité s'ajoute donc une baisse du taux de réponse d'où des risques potentiels de biais.

Comme les questions sont souvent posées à propos de sujets voisins, il est logique que les réponses obtenues soient proches et que les variables observées soient parfois fortement corrélées. Ceci entraine un problème de multi-colinéarité, qui a fait l'objet de nombreuses études dans l'industrie des études de marché.

Remarquons à ce sujet deux éléments contradictoires. Dans les études sociales et psychologiques, il est souvent fait recours à l'alpha de Cronbach, qui permet de justifier que des observations sont suffisamment cohérentes entre elles pour construire par exemple un indice. En fait plus les variables sont corrélées, plus l'alpha de Cronbach est élevé (plus il est en fait proche de 1) plus les variables mesurées sont colinéaires. Un « bon » indice de Cronbach est généralement recherché à hauteur de 0,7 voire 0,9. De ce point de vue il est donc souhaité que les variables observées soient corrélées, pour avoir une confirmation de la cohérence des observations et une confiance dans la pertinence de l'indice construit à partir de leur combinaison comme par exemple une moyenne. Dans les études par enquêtes, il est ainsi d'une certaine façon rassurant de constater que des variables sont cohérentes, participant à la mesure d'un phénomène donné comme par exemple l'attachement à une marque commerciale. (*)

Mais cette cohérence devient précisément un problème au moment de l'analyse car elle crée naturellement une colinéarité. Rechercher et être conforté par de bons alphas de Cronbach est donc contradictoire avec le souhait d'éviter la colinéarité. Notre point de vue dans cette recherche est précisément de rejeter l'idée que la colinéarité serait une sorte de « maladie » qu'il faut « traiter » ou « surmonter » (plusieurs publications expliquent ainsi « how to deal with » ou « how to overcome multicollinearity ». En fait il faut considérer que la colinéarité est en soi une information intéressante, qu'il faut intégrer de façon pertinente dans la conception même des modèles et méthodes d'analyse. En effet, le choix des questions renvoie à des concepts préalables issus de l'expertise du sujet quant aux phénomènes observés, ce qui permet de légitimer *a priori* certaines options de modélisation conformément à un modèle théorique.

(*) **Note** : *L'utilisation de l'α de Cronbach a fait l'objet de critiques. (Schmitt (1996)).*

Ceci conduit à des approches de modélisation postulant certains effets directs et indirects. Notons que les « avis d'expert » conduisant à utiliser un modèle conceptuel préexistant sont parfois intuitifs mais que dans la pratique

courante du Market research leur validité sous-jacente mériterait plus fréquemment une confrontation avec des méthodes de confirmation de modèles.

Au plan des données elles-mêmes, signalons que nous travaillerons ici la plupart du temps sur des jeux de données tels qu'obtenus à la fin du processus d'enquête et analyserons donc les résultats sans revenir jusqu'aux fondements mêmes c'est-à-dire sans traiter l'ensemble des questions spécifiques d'échantillonnage (comme le biais ou la couverture) ou de conception de l'enquête (comme le choix et l'ordre des questions dans le questionnaire).
Ainsi il est courant que l'ordre dans lequel des questions sont posées puisse influer sur les réponses obtenues. Un exemple caractéristique est appelé effet de halo, qui est l'effet obtenu dans l'ordre des questions en orientant le répondant dans un certain thème pour lui poser finalement une question spécifique. Si des individus sont d'abord interrogés sur leur opinion à propos de faits divers associés à des crimes choquants, ils répondront dans la suite du questionnaire plus sévèrement à des questions sur la politique pénale ou la peine capitale. Un autre exemple est fourni par les études de satisfaction, où la satisfaction globale est mesurée de façon plus stable en début de questionnaire plutôt qu'à la fin. En effet si les détails du questionnaire de satisfaction évoluent comme dans le cas de mesures continues au fil des années, les changements éventuels dans le détail du reste du questionnaire peuvent introduire une rupture dans la mesure de la satisfaction globale si celle-ci est notée à la fin de l'interview. Ces exemples illustrent le fait que les résultats obtenus peuvent dépendre de multiples facteurs qui resteront hors du champ de cette recherche.

Nous utiliserons cependant des propriétés pertinentes des variables et données, comme la taille d'échantillon qui permet de quantifier les précisions de certaines estimations, et tirerons parfois partie de la formulation connue des questions posées, qui peut jouer un rôle utile dans l'interprétation.

Dans les études de marché, les variables peuvent être de natures différentes. Ces variables sont communément classées en quatre types :

- Questions ouvertes : (pré ou post codées)
- Questions « fermées » avec trois sous- catégories :
  - nominales (les modalités servent au classement : par exemple sexe ou type de véhicule possédé)
  - ordinales (les modalités indiquent la position relative des objets entre eux) : elles montrent le choix du répondant et ses préférences entre différentes options
  - continues (échelles, ratios, intervalles).

Les variables nominales les plus systématiques sont les Catégories Socio Professionnelles, l'âge, le sexe, la région de résidence, la profession, etc…

Les variables ordinales sont souvent utilisées pour des techniques dites « trade-off » c'est-à-dire pour obtenir des répondants des classements entre produits et options et déterminer ceux qui ont le plus de chance d'entraîner un choix positif et ce résultat sera donc utilisé dans la conception de produits pour arbitrer entre différentes options de design et caractéristiques.

Les variables continues peuvent être un niveau de revenus ou de dépense, les variables ordinales peuvent être un niveau d'adhésion à une proposition formulée (ex : irez-vous votez « surement, peut-être etc…). Notons que les réponses graduées à une question donnée, comme les réponses sur une échelle de Likert (les répondants notant par exemple de 1 pour très mauvais à 10 pour excellent la satisfaction à un service) sont souvent traitées comme des variables continues bien qu'elles puissent être vues comme des variables ordinales.

Il peut aussi arriver que les données ne soient pas uniquement issues de l'enquête, mais proviennent d'autres sources ou bases de données. Ainsi dans le cas d'une enquête auprès de clients enregistrés dans un programme de fidélité, les réponses d'un individu peuvent être associées à des données d'achat. De la même façon si un passager d'une compagnie aérienne est interrogé, il est possible d'associer des données du système CRM (Customer Relationship Management) aux données d'enquête, pour analyser par exemple des impacts selon que le passager a réservé en ligne ou par téléphone, s'est enregistré lui-même ou à l'aéroport etc…

Dans le processus de réalisation de l'enquête une fois l'enquête terminée, les données font l'objet d'un prétraitement (appelé DP en anglais pour Data Processing) pour vérifier la cohérence des données éliminer les éventuelles aberrations, doublons etc., et le cas échéant introduire la pondération souhaitée (poids respectif des individus par exemple dans le cas de la méthode des quotas). C'est généralement à partir de ce fichier validé (clean file) que commence véritablement le processus d'analyse des données (tris croisés, analyse de la matrice des covariances ou équivalents) puis de modélisation (analyses multivariées, segmentations, régressions, prévisions).

Remarquons que la définition d'ESOMAR est assez large en ce qui concerne les outils d'analyse : *"using the statistical and analytical methods and techniques of the applied sciences to gain insight or support decision making"*. Elle couvre donc aussi le *Statistical Learning*, la simulation ou tout autre type de méthode issue des « sciences appliquées ».

1.2 L'analyse des leviers

Nous allons nous intéresser à certaines des catégories de méthodes référencées dans l'industrie des études de marchés, effectuer une analyse critique et proposer des recommandations. Nous explorerons aussi l'apport potentiel de méthodes issues d'autres disciplines. Soulignons finalement l'objectif clairement affiché par ESOMAR : *« to*

gain insight or support decision making ». Dans le cas de l'analyse des leviers cela signifie les perspectives suivantes :

Expliquer : ceci peut viser à identifier des facteurs particuliers comme des caractéristiques d'une marque ou des facteurs politiques ayant un impact sur un effet attendu, comme par exemple une intention d'achat ou une intention de vote. Mais il n'est pas forcément possible d'agir sur tous ces facteurs. Ainsi les attributs d'une marque comme par exemple la familiarité *(« treat me like a friend* »), ou la proximité *(« I feel close to that brand* ») peuvent être très explicatives tout comme la satisfaction à l'égard d'un élément de service (rapidité de prise en charge de mon véhicule au garage). Cependant agir sur les attributs de marque risque de prendre plus de temps et être moins aisé que d'agir sur des caractéristiques opérationnelles de la prestation de service. L'analyse des leviers est fréquemment utilisée en vue d'une explication des facteurs expliquant un effet pour fournir un critère de prise de décision. Notons qu'en réalité l'objectif d'explication peut être poursuivi en utilisant aussi bien des modèles d'exploration (Exploratory Models) que de confirmation (Confirmatory Models) afin de confirmer un modèle d'expert.

L'explication peut être présentée de plusieurs façons : soit en proposant une hiérarchisation des facteurs explicatifs (impact fort, moyen ou faible) soit en proposant une quantification des impacts par des valeurs calculées. Les résultats ne sont pas toujours présentés de façon claire quant à la nature exacte du modèle utilisé. En particulier certains chiffres peuvent représenter une « part d'impact total » mais ne pas être directement utilisables pour calculer un résultat qui serait obtenu sur l'objectif (par exemple en cas de décomposition de la variance, comme expliqué plus loin). En effet les utilisateurs sont souvent intéressés à quantifier un « what if » c'est-à-dire quel est l'impact quantifiable sur l'objectif recherché d'une action donnée sur les facteurs explications sur lesquels une action peut être engagée et mesurée.

Dans de nombreux cas l'analyse effectuée combinera une évaluation de l'importance avec une évaluation de la performance. Typiquement cette analyse (IPA en anglais pour Importance Performance Analysis) classera les leviers d'action, c'est-à-dire les prédicteurs dans un quadrant avec deux axes (Importance/Performance) et deux niveaux (Elevé, Faible). (Cf. Martilla et James (1977)).

La gestion des priorités s'en déduit :

- d'abord les leviers de faible performance mais de forte importance : fort gain potentiel.
- les leviers de forte importance et de performance élevée seront des éléments « à maintenir » : forte perte potentielle en cas de dégradation de la performance.
- viennent ensuite les leviers de faible performance et faible importance.

- la dernière catégorie (faible importance forte performance) ne nécessite en général aucune action particulière.

Simuler : ceci visera à analyser l'efficacité relative des leviers en simulant l'impact d'actions d'amélioration au niveau des variables levier, puis en calculant l'impact sur l'effet d'intérêt du modèle. Les techniques de simulation sont donc aussi un moyen de procéder à l'analyse des leviers avec un souci de quantification des impacts dans une approche explicative. Dans cette perspective la faisabilité concrète d'une amélioration au niveau des prédicteurs sera prise en compte, notamment si la performance de ce prédicteur est faible et qu'une amélioration est considérée comme opérationnellement faisable.

Prédire : l'objectif d'une prédiction dans les études de marché est plus compliqué. En effet une fois le plan d'action mis en œuvre, le critère objectif pourra être de nouveau mesuré et comparé à la prédiction antérieure. Mais si le modèle n'explique qu'une partie assez limitée du phénomène observé, la prévision risque d'être peu fiable. Ainsi par exemple si le modèle n'inclut pas l'impact des actions de la concurrence, la situation peut avoir évolué sur le marché considéré et le modèle par définition ne pourra pas le prendre en compte.

Même dans l'hypothèse où les données permettraient en théorie une modélisation suffisamment prédictive, la notion de modèle renvoie communément à une représentation de la réalité mais qui peut offrir une variété de techniques et de choix de prévision. A l'extrême certains modèles performants ne peuvent pas être interprétés. Même dans le cas de modèles simples avec des hypothèses claires et explicites de causalité, il est possible de modéliser analytiquement de plusieurs façons. Par exemple choisir dans une régression ou dans des modèles structurels de considérer que chaque prédicteur peut évoluer isolément, ou au contraire prendre en compte des évolutions liées de plusieurs prédicteurs pour tenir compte de certaines associations entre les facteurs.

L'analyse des leviers est ainsi mise en œuvre de façon variée mais la référence de base dans de nombreux domaines d'application comme le marketing, la psychologie ou la sociologie est la régression linéaire multiple et la quantification des leviers par les coefficients β standardisés. Cependant cette approche présente des écueils dans la plupart des cas pratiques en raison des corrélations entre prédicteurs. Depuis les années 1980 en raison de l'accroissement des capacités de calcul informatique, sont apparues différentes méthodes en particulier celles fondées sur la décomposition de la variance en vue de générer des importances relatives toutes positives et plus stables. Plusieurs méthodes seront étudiées plus loin.
Mais de fait la référence de base mise en avant reste la régression linéaire (OLS). A titre d'exemple dans une documentation technique fournie par IBM SPSS et disponible en ligne sur le site de l'entreprise ainsi que sur d'autres sites (comme www.kdnuggets.com) intitulée : « How to Get More Value from your Survey Data » (cf. bibliographie) la régression linéaire simple est présentée comme *« the most popular method for studying the relationship between an outcome variable and several predictors or independent variables »*. Relevons qu'IBM SPSS utilise le terme répandu dans de *« predictor, or independent variables »* que nous proscrirons ici. Nous

utiliserons les termes « variable à prédire » et « variables explicatives ou prédicteurs ». Ce document ajoute que bien que les relations entre les variables peuvent ne pas être linéaires, il est *« best to assume that they are in order to create a useful model* ». Plus loin il est également précisé que les meilleurs prédicteurs devraient de préférence être sélectionnés pendant l'analyse, car bien que le modèle puisse être utilisé avec un grand nombre de prédicteurs, il est préférable d'être « plus sélectif ». Enfin, l'importance des prédicteurs est quantifiée via les coefficients standardisés.

Cette utilisation de la régression multiple par IBM SPSS est intéressante car d'une part elle émane d'un acteur de référence dans le domaine des outils statistiques fournis aux sociétés d'étude de marché, mais d'autre part elle omet plusieurs des difficultés principales rencontrées par les opérateurs d'études de marché.

La première question totalement évacuée est la non linéarité. Ainsi la « courbe de réponse » des variations de la variable modélisée par rapport aux scores de performance sur différents prédicteurs peut ne pas être convenablement représentée par une droite.

Certains prédicteurs sont ainsi impactants en cas de faible performance mais ne jouent plus guère de rôle au-delà d'un certain niveau : il s'agit des prérequis. (par exemple la lisibilité de la facturation qui ne joue pas un rôle de satisfaction accrue si elle est imprimée sur papier glacé). A l'inverse d'autres pourront ne pas avoir d'impact négatif en cas d'absence mais au contraire jouer un rôle de satisfaction accrue en cas de mise en œuvre. Ce sera le cas pour des facteurs de contentement appelés communément « delight » (par exemple une peluche cadeau pour les enfants dans une chambre d'hôtel). Enfin certains leviers seront opérants des deux côtés, jouant un rôle à tous les niveaux de performance.

La deuxième question évacuée est précisément celle de la multi-colinéarité qui est traitée en recommandant de limiter le nombre des questions et de choisir de façon pertinente les prédicteurs. Si dans le cas présenté par IBM SPSS dans la documentation citée plus haut et référencée en bibliographie l'usage direct de la régression multiple fonctionne assez bien, dans les cas courants les difficultés rencontrées (non linéarité, nombreux prédicteurs, multi-colinéarité) induisent des problèmes dans les calculs et l'interprétation.
Ceci a poussé les entreprises du secteur des études de marché à rechercher des solutions alternatives à la régression linéaire pour estimer l'importance des prédicteurs.

L'analyse des leviers et de la causalité est également abordée dans l'ouvrage de Giannelloni et Vernette précité. Dans un premier temps, au chapitre 13 intitulé « Analyse d'association et de causalité », l'ouvrage explique comment étudier et aborder la question de l'association entre variables, de l'intensité de cette relation et des éventuels sens de causalité de cette relation. De la même façon, le chapitre traite de la sélection du type de modèle d'ajustement et de la qualité d'ajustement des observations. Néanmoins, dans ce chapitre, qu'il s'agisse de l'analyse du coefficient de détermination et de sa significativité ou bien de l'utilisation pour le cas de l'association entre

variables non métriques, de la corrélation des rangs de Spearman ou bien du *tau* de Kendall, il s'agit dans tous les cas de simples analyses des techniques et de la validité des éléments de mesure d'une force et d'un sens d'association entre variables mais cela ne porte pas sur une réflexion et une analyse de la causalité en elle-même.

Le sujet est néanmoins de nouveau évoqué dans le chapitre 15, dont le titre est *« L'expérimentation en marketing : collecte et traitement des données ».*
Il y a là clairement une présentation dans ce chapitre du but qui consiste à chercher à identifier des liens de causes à effets. Après un développement sur les concepts de cause et de causalité, l'ouvrage met en avant trois manifestations empiriques de la causalité :

- La variation concomitante. Ceci renvoie au Chapitre 13 précédent dans l'ouvrage sur la nécessité d'avoir une corrélation, une association, entre une cause et un effet.
- L'ordre de manifestation des variables dans le temps. L'ouvrage souligne qu'en bonne logique, une cause doit précéder un effet.
- L'élimination des autres sources possibles de cause.

A condition donc, poursuit l'ouvrage, de mettre en place avec rigueur ces principes, les méthodes expérimentales sont recommandées pour identifier les causes à effets. Notons que les trois manifestations empiriques de la causalité précitées sont totalement identiques et dans le même ordre que celles retenues par Terry Grapentine (2012).

Aussi, cet ouvrage n'aborde pas en tant que telle la recherche des causalités ou l'inférence des causalités à partir des données et il n'y a pas de référence explicite aux réseaux bayésiens et aux approches de l'analyse de la causalité comme développée par exemple par Judea Pearl (2009).

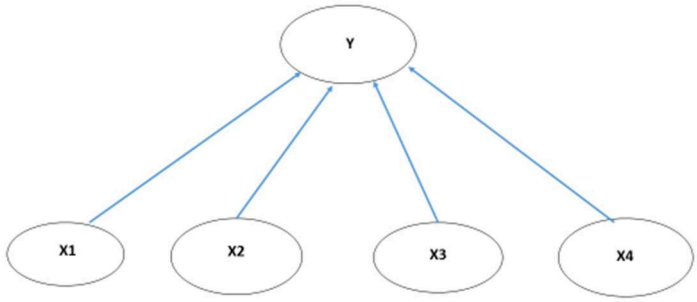
L'analyse des leviers est également abordée dans l'ouvrage au chapitre traitant de la régression multiple. Il y est souligné que les variables « indépendantes » devraient l'être réellement, et que dans le cas contraire apparait un effet de multi-colinéarité dont les conséquences sont doubles : estimation imprécise des coefficients et manque de stabilité, c'est-à-dire que les valeurs des coefficients varieront beaucoup d'un échantillon à l'autre. L'ouvrage recommande donc *« chaque fois que c'est possible, de privilégier l'indépendance dans le choix des variables explicatives potentielles ».* Comme indiqué plus haut même si cette recommandation est parfaitement légitime elle reflète un point de vue théorique qui est fort éloigné des situations couramment rencontrées dans la pratique.
Il apparaît donc que dans deux publications de base utilisées pour l'enseignement ou la formation professionnelle, la méthode de régression linéaire est présentée comme la plus courante avec recommandation d'éviter par design au départ dans le choix ou en fait plus implicitement par sélection des variables ou encore utilisation de composantes factorielles, de tenter d'éviter de se trouver en présence d'un effet de multi-colinéarité.

Aucune de ces références ne propose d'emblée de pistes pour réaliser l'analyse en cas de non linéarité ou de multi-colinéarité, tant en termes de diagnostics que de méthodes de traitement. Les praticiens d'études de marchés ne paraissent donc pas disposer dans la littérature technique de référence générale et communément admise sur l'analyse des leviers dans les études de marché. En revanche il existe une variété de publications par des individus ou des entreprises qui promeuvent des options techniques de façons plus ou moins détaillées, options qui ont eu leur nécessité vu l'absence de référence communément admise. Les méthodes proposées peuvent être classées en deux grandes catégories, selon que le praticien utilise ou non une approche structurelle dans la modélisation.

Détaillons les deux catégories :

Sans approche structurelle.

Les prédicteurs sont utilisés directement pour modéliser la variable à prédire. Par exemple la variable à prédire sera régressée sur l'ensemble des prédicteurs. Les approches de régularisation et de décomposition de la variance entrent aussi dans ce cadre. Dans les modèles directs, la variable à prédire est modélisée directement par l'ensemble des leviers comme indiqué ci-dessous.



Le modèle de régression où la variable à prédire est régressée directement sur les prédicteurs est un exemple de ce type de modèle.

Alternativement nous pouvons classer dans une deuxième catégorie les modéles utilisant une structure entre les variables, tels les modéles de chemin (path models) ou encore les réseaux bayesiens.

Avec approche structurelle.

Le modèle va incorporer des relations intermédiaires entre les prédicteurs. Ceci est pratiqué de différentes façons :

- Certaines variables mesurées sont considérées comme « prédicteurs clés » (« Key Metrics »). Elles vont donc faire l'objet d'une question dans le questionnaire, et seront supposées souvent résumer l'information acquise avec d'autres questions. Par exemple la satisfaction avec l'accueil en agence sera explicitement demandée, et sera une « key metric » utilisée dans la modélisation de la satisfaction globale, tandis que des questions plus détaillées relatives à la satisfaction avec l'accueil en agence seront également demandées (relatives par exemple au temps d'attente, à la satisfaction avec la diligence du personnel etc..). Une modélisation structurelle sera donc réalisée selon un plan analytique qui sera décidé par le praticien.

- Des facteurs sont identifiés et utilisés comme prédicteurs intermédiaires.

- Utilisation de variables latentes : par distinction avec les variables manifestes ou observables, les variables latentes ne sont pas directement observables mais sont liées aux variables observables tout en étant considérées comme plus représentatives d'un concept comme par exemple l'attachement à une marque. Dans ce cas le praticien va choisir différentes variables latentes.

Ces approches sont utilisées dans le cadre des modèles d'équations structurelles qui sont issus des recherches sur la causalité. Les modèles d'équations structurelles ont été mis en œuvre dans le domaine des études de marché particulièrement pour les études de satisfaction, mais aussi en psychologie et sociologie. Deux méthodes sont particulièrement utilisées : l'approche LISREL (Jöreskog et al., (1982)) et PLS (Wold, (1975 ;1985)).
Il existe en fait une grande variété de pratiques avec une terminologie non standardisée, consistant à appliquer des méthodes statistiques en vue de quantifier des relations de cause à effet résultant d'un modèle théorique, avec la possibilité de prendre en compte des concepts latents mesurés à partir de variables observables. (Bollen, (1989) ; Kaplan, (2000)).

Enfin ont été développées depuis une dizaine d'années des techniques de Machine Learning telles les forêts aléatoires, ou les réseaux bayésiens.

Les prestataires d'étude de marché mettent ainsi en œuvre de nombreuses variantes pour deux principales raisons : soit pour traiter différentes difficultés rencontrées dans l'usage de la régression linéaire, soit dans le souci de développer une approche marketing de différenciation ou d'innovation par rapport à leur marché.

Ainsi dans les dix dernières années, les entreprises d'études de marché ont commencé à utiliser et présenter des méthodes d'analyse des leviers fondées sur la décomposition de la variance qui ont été présentées par plusieurs auteurs comme un moyen de traiter les conséquences de la multi-colinéarité. L'usage de la décomposition de la variance comme méthode de quantification de l'importance des prédicteurs a été également promu dans d'autres domaines comme dans plusieurs champs de la psychologie ou de l'économie. Tonidandel et al. (2011) font une place aux méthodes de décomposition de la variance.

Nous présenterons d'abord la problématique théorique de l'importance relative des prédicteurs et ensuite les méthodes de décomposition de la variance qui peuvent être considérées comme des cas particulier de quantification de l'importance relative.

Dans le cas particulier des études de marché les méthodes de décompositions de la variance représentent une tendance récente gagnant en visibilité depuis le milieu des années 2000. Cependant il paraît utile avant d'analyser des méthodes de quantification de l'importance de mieux expliciter le concept même d'importance relative entre prédicteurs.

Chapitre 2 : L'importance relative

La définition de l'importance relative n'est pratiquement jamais explicitée dans la présentation des résultats d'études. En réalité l'emploi du terme « importance » est source de confusion, et a été abordée de plusieurs façons et par différents auteurs (Johnson (2000), Cosnefroy et Sabatier (2011)). Nous allons présenter ci-après le contexte et des références sur ce sujet.

2.1 Importance relative des prédicteurs : contexte et références.

La comparaison des prédicteurs en termes d'importance relative a fait l'objet de différentes solutions proposées depuis 1960 jusqu'aux années 2000. Tout d'abord, l'approche sera différente dans l'exploitation des modèles directs (c'est à dire non structurels) de régression selon qu'il s'agira d'une perspective de prévision ou d'explication.

Dans le cas d'une approche de prévision l'apport collectif combiné des prédicteurs (le R²) sera typiquement la valeur intéressante plutôt que de savoir distinguer la contribution relative des prédicteurs. Nous sommes là dans l'approche qui consiste à se satisfaire d'une prévision efficace sans pour autant chercher à comprendre en détail les contributions relatives des prédicteurs.

Dans le cas d'une approche explicative, des chercheurs qui se sont intéressés à la notion d'importance ont identifié deux étapes majeures dans l'utilisation des modèles de régression : sélection des variables du modèle et puis comparaison des prédicteurs. *(*Cf. Cosnefroy Sabatier (2011) ; Azen et Budescu, (2003), Azen (2003).

En ce sens les prédicteurs abandonnés au stade de la sélection sont réputés d'importance nulle. Par exemple dans certains cas les clients voudront identifier les variables d'importance en cherchant un sous ensemble de prédicteurs qui ensemble conduisent à un R² aussi élevé que possible. Pour minimiser l'erreur de prévision d'autres critères sont cependant recommandés comme $\hat{\sigma}^2 = (\frac{n}{n-k-1})(1-R^2)s_y^2$, ou le Press. Si le nombre de prédicteurs à considérer est imposé, le R^2 maximum sera recherché, si p n'est pas imposé $\hat{\sigma}$ sera minimisé. Des méthodes de sélection pas à pas (descendantes ou ascendantes, stepwise) sont notamment disponibles.

Schafer (1991) puis Nathan, Oswald, Nimon (2012)) ont considéré des méthodes où l'ordre des prédicteurs tient compte d'une hiérarchie de pertinence connue *a priori* (relevant known ordering) et se placent donc dans un processus ou chaque variable est introduite en fonction d'une théorie préalable.

Dans le cas où un modèle organisé est postulé, par exemple si des variables ont des effets directs et indirects, le modèle analytique pourra être un modèle de chemin (path model). Nous n'analyserons pas dans cette section la

notion d'importance dans le cas d'un modèle structuré ou ordonné car elle est naturellement différente, nous nous intéressons donc au modèle direct sans *a priori* sur la pertinence des prédicteurs.

Achen (1982) classifie trois différentes catégories d'approches quant à la quantification de l'importance relative des prédicteurs :

Importance Théorique (Theoretical Importance) : c'est le lien entre entre un changement du critère associé à un changement sur le prédicteur. Ceci peut être formalisé de très différentes façons. Considérons le cas où nous disposons de n observations d'une variable y et de p prédicteurs $x_1,...x_p$. Nous disposons aussi d'une fonction f qui représente la dépendance sur l'espérance conditionnelle de y sachant $(x_1,...,x_p)$. Nous pouvons associer à chaque observation k des valeurs des prédicteurs notées $x_{k1},...x_{kp}$ une valeur modélisée $\hat{y}_k$:

$$\hat{y}_k = \mathrm{f}(x_{k1},...x_{kp})$$

Pour quantifier l'impact du prédicteur j il est par exemple possible de simuler un accroissement aléatoire pour chaque observation k du prédicteur j δx_{kj}.

Soient :

$$\delta y_{k,j} = \mathrm{f}(x_{k1},..x_{kj} + \delta x_{kj},....x_{kp}) - \mathrm{f}(x_{k1},..x_{kj},...x_{kp})$$

$$\Delta_{yj} = \frac{1}{n}(\sum_{k=1}^{k=n} \delta y_{kj})$$

$$\Delta_{xj} = \frac{1}{n}(\sum_{k=1}^{k=n} \delta x_{kj})$$

Alors l'importance simulée peut être quantifiée comme :

$$\hat{I}_{x_j} = \frac{\Delta_{yj}}{\Delta_{xj}}$$

Cette approche sera reprise au 3.1.11 dans la partie consacrée à l'analyse de sensibilité.

Dans le cas du modèle linéaire $\hat{I}(\mathbf{x_j})$ égale au coefficient non standardisé de la régression multiple. Dans le cas du modèle logistique ce résultat sera plus complexe à quantifier en raison de la non-linéarité. Il y a bien sur de multiples autres manières de calculer une « importance théorique » au sens d'Achen, en utilisant différents types de modèles (logistique, linéaire...), selon la méthode de simulation choisie (par exemple accroissement aléatoire ou non) ou encore selon les hypothèses de corrélation effective entre les prédicteurs c'est-à-dire si par exemple on admet qu'ils peuvent évoluer indépendamment ou qu'une action sur un prédicteur est associée à des évolutions sur d'autres prédicteurs. L'importance théorique est le résultat d'un ensemble d'hypothèses de modélisation, d'où une extrême variété de calculs possibles.

Contribution en niveau (Level Importance) : c'est la contribution du prédicteur à l'accroissement en niveau de la moyenne de la variable à prédire, donc en fait dans le cas du modèle linéaire c'est le produit de la moyenne observée du prédicteur par le coefficient de régression non standardisé. Cette contribution en niveau est répandue en économie (Johnson (2004), Kruskal & Majors, 1989). Elle revient alors à considérer la variable à prédire comme la somme de contributions indépendantes. Notons que cette approche n'est pas naturellement compatible avec des modèles du type logistique.

Part de variance (Dispersion Importance) : ceci renvoie à l'attribution à chaque prédicteur d'une part de variance expliquée. Cette mesure est fréquente dans les sciences du comportement dans le cadre de modèles explicatifs (Thomas et Decady 1999, Johnson & Lebreton, 2004). Nous reviendrons plus loin sur les différentes approches de décomposition de la variance qui ont été progressivement introduites dans les études de marchés dans les années 2000.

Devant cette variété de définitions de l'importance, mais aussi en raison de la complexité apparente d'une modélisation, les praticiens ont tenté d'utiliser directement la « zero order correlation », c'est-à-dire le simple coefficient de corrélation bivariée entre la variable à prédire et un prédicteur donné. Quand les prédicteurs sont deux à deux décorrélés le coefficient de corrélation bi varié entre le prédicteur et la variable à prédire est certes égal au β_i standardisé. Mais dès l'apparition de corrélations entre prédicteurs ce résultat n'est évidemment plus vrai : $\beta_j \neq r(y, x_j)$ pour au moins un des prédicteurs.

Cette quantification de l'importance par la corrélation bivariée a bien sûr été considérée depuis longtemps comme inadéquate (Hoffman (1960) ; Green & Tull (1975) ; Budescu (1993)), mais elle persiste dans la pratique en raison de sa simplicité, même si elle contredit l'intérêt d'une approche multivariée et exclut la prise en compte des interactions possibles.

En ce qui concerne la comparaison des prédicteurs en termes d'importance relative ce sujet a été étudié avec plusieurs solutions proposées depuis 1960 jusqu'aux années 2000. (Cf. le résumé historique de Johnson et Lebreton (2004)). L'importance relative avait certes été un sujet d'intérêt depuis longtemps pour les chercheurs (cf. Engelhart (1936) cité par Johnson et Lebreton (2004)), mais a pris un tournant dans les années 1960 avec les travaux d'Hoffman qui cherchait à décrire les processus cognitifs mis en œuvre par les cliniciens quand ils formaient des jugements sur le cas de leurs patients.

Hoffman utilisa le fait que la variance expliquée peut s'écrire comme la somme des produits du coefficient standardisé de la régression multiple multiplié par le coefficient de corrélation simple entre ce prédicteur standardisé et la variable à prédire, ainsi que présenté ci-après dans le cas de la régression linéaire.

Considérons que nous avons observé sur n individus $p+1$ variables représentées par des vecteurs de $\mathbf{R}^n$ $\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_p}$, $\mathbf{y}$ est la variable à prédire et les $\mathbf{x_j}$ sont les prédicteurs, nous pouvons aussi écrire :

$$R^2 = \frac{V(\mathbf{y}^*)}{V(\mathbf{y})} = \frac{\text{cov}(\mathbf{y}, \mathbf{y}^*)}{V(\mathbf{y})} = \frac{\sum_{j=1}^{j=p} b_j \text{ cov}(\mathbf{y}, \mathbf{x_j})}{V(\mathbf{y})} = \sum_{j=1}^{j=p} b_j \frac{s_{x_j}}{s_y} r(\mathbf{y}, \mathbf{x_j})$$

Soit encore en utilisant les coefficients de régression standardisés $\beta_j = b_j \frac{s_{x_j}}{s_y}$:

$$R^2 = \sum_{j=1}^{j=p} \beta_j \, r(\mathbf{y}, \mathbf{x_j}) \tag{2.1}$$

Nous reviendrons sur cette approche plus loin dans la partie consacrée à la décomposition de la variance. Hoffman proposa d'utiliser cette décomposition de la variance expliquée comme une quantification de la « contribution indépendante » de chaque prédicteur. Naturellement le choix même du mot indépendant dans ce contexte est confus, car il peut exister une corrélation non nulle entre ce prédicteur et les autres. Il aurait été sans doute meilleur de qualifier cette valeur de contribution « marginale ». Notons que Hoffman soutenait l'emploi du terme « indépendante » au motif que chaque terme $\beta_j \, r(\mathbf{y}, \mathbf{x_j})$ correspondait à un accroissement marginal de la covariance entre les valeurs prédites et le résultat d'un accroissement marginal des valeurs d'un prédicteur, les autres étant maintenus constants.

En réalité l'une des raisons invoquées parfois pour rejeter l'importance théorique est précisément (Lebreton (2004)) que dans le modèle conceptuel sous-jacent le praticien ne va pas considérer qu'un prédicteur peut évoluer les autres étant maintenus constants, mais au contraire que plusieurs vont évoluer ensemble. Notons que ceci peut être réalisé en ayant recours à un modèle structurel pour refléter la conviction que les prédicteurs évoluent de façon liée.

Comme ni le coefficient de corrélation entre la variable à prédire et un prédicteur, ni le coefficient de régression ne permettait d'obtenir une réponse satisfaisante dans les perspectives mentionnées ci-dessus, Courville et Thomson (2001) ont proposé d'analyser l'importance en tenant compte simultanément du coefficient de régression et du coefficient de corrélation bivariée entre chaque prédicteur et la variable à prédire. Cette analyse était faite en considérant que l'usage du β standardisé sans une analyse de la structure des prédicteurs pouvait conduire à des interprétations erronées.

Cependant utiliser deux indices au lieu d'un seul rend plus complexe l'attribution d'une importance. Il y a donc eu au fil des années des tentatives variées pour proposer une méthode convaincante de quantification de l'importance. Après Hoffman, Gibson (1962) suggéra d'utiliser des facteurs orthogonaux « proches » des variables de départ. Puis Darlington (1968) passa en revue plusieurs méthodes, critiqua la décomposition de Hoffman et conclut qu'aucune méthode ne permettait de résoudre complètement la question dès qu'il y avait corrélation entre prédicteurs.

Les limitations de la régression multiple ayant été mises en avant par plusieurs auteurs tels Green & Tull (1975), McLaughlan (1992), puis Kruskal et Majors (1989), Johnson et Lebreton (2004) ont examiné le concept d'importance dans différentes domaines (psychologie, économie, physique) et confirmèrent que cette question intéressait bel et bien plusieurs disciplines sans pour autant proposer de solution de référence.

Aussi certains travaux ont conduit à des propositions identiques malgré des cheminements différents. Ainsi en 1980 Lindemann, Merenda et Gold proposèrent d'utiliser des séquences de R^2 obtenues en moyennant les R^2 obtenus sur les sous-ensembles ordonnés de prédicteurs (méthodes dite *lmg* tirée du nom des trois auteurs, ceci sera aussi revu dans le chapitre consacré à la décomposition de la variance). Azen (2003) et Budescu (1993) introduisirent l'analyse de dominance qui en pratique donne les mêmes mesures que LMG. Conklin et Lipovetsky (2004) réinventèrent le concept en utilisant la Shapley Value.

Pour plus de détails sur ces perspectives historiques voir Lebreton (2004) et Grömping (2007). L'approche des poids relatifs (Relative Weights Analysis) a également été proposée avec des résultats identiques entre Genizi (1993) puis Johnson (2000), ce qui constitue un autre exemple de convergence. Il sera cependant établi plus loin que ces deux méthodes ne sont pas identiques à la méthode proposée par Fabbris (1980) contrairement à ce qui était indiqué par Grömping (2015).

2.2 Importance des prédicteurs : formalisation et méthodes utilisées.

La problématique d'importance consiste à allouer à chaque prédicteur une valeur numérique représentant son « importance » par rapport à un y . De façon intuitive l'idée est qu'une modification d'une des variables X_j jouant un rôle de prédicteur influencera les valeurs prises par la variable à prédire y .

Il existe donc une grande variété de modèles possibles. Nous pouvons définir une fonction d'importance de la façon suivante :

Soit un ensemble de p variables aléatoires réelles X_j constituant un ensemble de prédicteurs.

Notons P = {1,…,p} le sous-ensemble composé des indices de ces p variables.
Soit y une variable à prédire également sous la forme de variable aléatoire réelle.

Définition : Fonction d'Importance :

Une fonction d'importance est une fonction : $I : P \rightarrow \mathbb{R}^p$ associant à chaque prédicteur (représenté par son indice) une valeur d'importance.

Au niveau de chaque prédicteur nous pouvons donc écrire :

$$I(X_j) = I(j, X, y)$$

Cette définition est très générale, et dans la pratique les auteurs ont considéré des fonctions d'importance en relation avec un modèle paramétrique (en général le modèle linéaire) et aussi en faisant dans certains cas (par exemple *lmg-Shapley* qui sera présentée plus loin)) intervenir des sous-modèles associés à des sous-ensembles des prédicteurs. Dans ce qui suivra nous proposerons des formulations diverses de l'importance théorique en utilisant le cadre d'analyse du modèle linéaire simple, et en considérant différents types de fonctions d'importance. Les fonctions d'importance proposées au plan théorique sont utilisées en pratique avec des jeux de données et en réalité les valeurs calculées sont des estimateurs. Ce point n'est en général pas relevé dans les articles. Seule Grömping (2007) fait explicitement référence à la notion d'estimateurs d'importance relative, ce qui est en réalité une distinction pertinente car dans la décomposition complète de la variance expliquée, ces estimateurs sont nécessairement biaisés. Grömping (2006, 2015) distingue une importance « conditionnelle » c'est-à-dire en conjonction avec les autres prédicteurs et une importance « marginale » c'est-à-dire individuelle comme par exemple le carré du coefficient de corrélation bivariée.

Dans la suite de ce travail nous nous intéresserons sauf exception explicite aux estimateurs de l'importance. Grömping (2006) a présenté différentes méthodes de quantification d'importance relative et s'est efforcée de les illustrer dans un document de synthèse en liaison avec la mise à disposition du package R `relaimpo`, qui est celui qui regroupe aujourd'hui le plus de méthodes dans un même package. Elle a identifié 6 estimateurs dans la première version de `relaimpo` (Grömping (2006)) et a en 2010 ajouté deux autres méthodes (*johnson* et *CAR*). Dans le cadre de cette recherche nous avons écrit 2 scripts additionnels, pour les méthodes de *green* et *fabbris* en corrigeant le code mentionné par Grömping (2015) pour la méthode *green* et en distinguant, contrairement à l'analyse de Grömping la méthode proposée par Fabbris (1980), en effet elle donne des résultats différents de la méthode de Johnson. Aussi il est présenté ici un nouveau script pour une variante dite *DCP* (Décomposition via Composantes Principales) puis aussi pour *weifila*, nouvelle méthode proposée dans cette recherche.

Après avoir examiné les 6 méthodes proposées initialement par Grömping dans la version originale de `relaimpo` nous étudierons les méthodes *johnson* et *CAR*, puis présenterons quatre approches : *green*, *fabbris*, *DCP* et *weifila* avec leurs scripts R, soit 12 méthodes au total. Pour 3 méthodes (*fabbris, green* et *CAR*) les résultats obtenus dans le cadre de cette recherche sont différents de certains résultats publiés à ce jour.
Les 6 premières méthodes présentées par Grömping sont :

- *first*
- *last*

- *beta square*
- *pratt*
- *lmg-Shapley*
- *pmvd*

Elles seront présentées en détail au chapitre 3. Grömping les présente avec le jeu de données *swiss* et formule des commentaires relatifs à chaque méthode de quantification de l'importance. Les résultats ainsi obtenus avec le jeu *swiss* seront reproduits ci-après dans un but simplement illustratif. Le jeu *swiss* est directement disponible dans la console R (cf. Grömping (2006)). Ce jeu de données comporte 47 observations (sur des provinces francophones) avec comme variable à prédire Fertility et les prédicteurs suivants :

- Agriculture (% hommes y travaillant)
- Examination (% Notes élevées à l'examen militaire)
- Education (% de conscrits ayant été au-delà de l'école primaire)
- Catholic (% de catholiques)
- Infant.Mortality (% de décès avant un an)

Dans la perspective de l'analyse des différentes quantifications de l'importance, ces données présentent l'intérêt d'avoir une certaine colinéarité entre plusieurs variables comme le montre la matrice de corrélation ci-dessous :

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	1,0000	0,3531	-0,6459	-0,6638	0,4637	0,4166
Agriculture	0,3531	1,0000	-0,6865	-0,6395	0,4011	-0,0609
Examination	-0,6459	-0,6865	1,0000	0,6984	-0,5727	-0,1140
Education	-0,6638	-0,6395	0,6984	1,0000	-0,1539	-0,0993
Catholic	0,4637	0,4011	-0,5727	-0,1539	1,0000	0,1755
Infant.Mortality	0,4166	-0,0609	-0,1140	-0,0993	0,1755	1,0000

Tableau 2.2.1 : Matrice des corrélations. Données *swiss*.

L'application directe de l'OLS donne les résultats suivants :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66,915180	10,706040	6,250000	0,000000
Agriculture	-0,172110	0,070300	-2,448000	0,018730
Examination	-0,258010	0,253880	-1,016000	0,315460
Education	-0,870940	0,183030	-4,758000	0,000024
Catholic	0,104120	0,035260	2,953000	0,005190
Infant.Mortality	1,077050	0,381720	2,822000	0,007340

Tableau 2.2.2 : Régression multiple. Variable à prédire : Fertility. Données : *swiss*.

Dans ce modèle le caractère non significatif du coefficient « Examination » ne surprend pas vu la nature des données. Le R^2 est de 0,7067.

2.2.1 Carré des corrélations bivariées (*first).*

Cette méthode consiste à affecter à chaque prédicteur une valeur d'importance proportionnelle au carré de la corrélation bivariée entre le prédicteur et la variable à prédire. Cette méthode a été qualifiée de « first » : l'emploi du terme « first » signifie simplement qu'ils sont entrés en premier et en fait seulement inclus dans chaque modèle, par comparaison aux autres choix et qui mettent en jeu des séquences possibles d'introduction de plusieurs prédicteurs dans le modèle.

Définition : Importance *first*

L'importance « first » du prédicteur j est :

$$I_{first}(X_j) = r(y, X_j) \tag{2.2.1.1}$$

Gromping souligne que la méthode *first* conduit souvent à une répartition dont la somme excédera fortement le R^2 du modèle complet : *"If regressors are correlated, the sum of these individual contributions is often far higher than the overall R² of the model with all regressors together, i.e., the overall model explains less than the sum of all individual models"*.(Gromping , (2006))

Dans le cas des données *swiss* les importances relatives *first* sont :

Agriculture	Examination	Education	Catholic	Infant.Mortality
0,1246649	0,4171645	0,4406156	0,2150035	0,1735189

Tableau 2.2.1.1. Importances *first*. Données *swiss*.

Et effectivement la somme des importances *first* est de 1,37097, soit environ le double du R^2.

Johnson et Lebreton (2004) entre autres ont critiqué cette méthode en raison de l'impasse faite sur les effets indirects potentiels entre drivers. Cela constitue en effet une approche excluant toute analyse des effets combinés des variables, donc en fait une approche univariée par comparaison avec une approche multivariée comme dans le cas

de la régression multiple. Mais relevons que même dans le cas de la régression multiple, l'utilisation des β_j ou de mesures dérivées des β_j peut être contestée selon les hypothèses choisies pour le modèle d'interaction entre prédicteurs. Bien que l'analyse des corrélations entre chaque prédicteur et la variable à prédire puisse constituer une étape utile dans l'analyse exploratoire des données, pour les mêmes raisons que les auteurs précités c'est-à-dire en particulier pour l'absence totale de prise en compte des interactions entre les prédicteurs, nous la considérons également comme à proscrire dans l'analyse de l'importance des leviers dans le cas des études de marchés.

2.2.2 Importance *last.*

Une autre approche de la quantification de l'importance relative consiste à l'inverse à quantifier la part de variance expliquée ajoutée « en dernier » par un prédicteur j par rapport au modèle constitué par les $p-1$ autres prédicteurs d'où le nom de *last*.

Définition : Importance *last*

Soit P l'ensemble des prédicteurs et P\{j} l'ensemble des p-1 prédicteurs excluant le prédicteur j.

$$I_{last}(X_j) = R^2(P) - R^2(P \setminus \{j\}) \qquad (2.2.2.1)$$

Les estimateurs associés s'en déduisent. Notons que cette définition de l'importance peut aussi être formalisée comme la contribution marginale d'un prédicteur considéré comme un joueur dans un jeu coopératif ou la fonction de gain est la variance expliquée par les prédicteurs dans le modèle linéaire simple. Nous reviendrons sur ce point à propos de l'utilisation de la Shapley Value et de la Owen value.

Ces parts de variance allouées sont identiques pour les utilisateurs de SAS aux Types II SS divisés par $\sum (y - \bar{y})^2$

. Cette métrique *last* a aussi été appelée "*usefulness*" par Darlington (1968).

Voici les résultats de la méthode *last* avec les données *swiss* :

Agriculture	Examination	Education	Catholic	Infant.Mortality
0,04286961	0,00738742	0,16196269	0,06237263	0,056945259

Tableau 2.2.2.1. Importances *last.* Données *swiss*.

Le résultat avec le jeu de données *swiss* est à mettre en liaison avec le caractère non significatif du coefficient de l'OLS pour le prédicteur « Examination ».

Par rapport aux résultats de la méthode *first,* les valeurs de la méthode *last* sont ici inférieures pour chaque variable. A ce propos, Grömping indique que si les régresseurs sont corrélés alors la somme des valeurs *last* pour l'ensemble des prédicteurs n'est pas égale à R^2 , mais sera « *typiquement* » très inférieure au R^2 . La variance n'est donc pas complètement décomposée. Aussi les effets indirects ne sont pas pris en compte d'où le rejet de ce type de méthode

par Johnson et Lebreton. Dans le cas des données *swiss*, la somme des valeurs *last* est 0,33154 effectivement inférieure à la variance expliquée qui est de 0,70674.

Toutefois, la somme des *last* n'est pas systématiquement inférieure au R^2. Ainsi Tabachnick et Fidell ((2001), Chapitre 5, 5.6.1.1) confirment comme Grömping que la somme des carrés des coefficients de corrélation semi-partiels est « *usuellement* » inférieure, mais en revanche ajoutent une précision en notant qu'il peut se produire que la somme des *last* soit supérieure au R². Tabachnick a indiqué séparément dans un échange de courrier que cela pouvait être le cas dans des configurations très particulières de la structure des prédicteurs.

Nous établirons plus loin au moment de l'étude de la décomposition de la variance que dans le cas de deux prédicteurs cette configuration sera rencontrée dans les cas où les deux conditions suivantes sont simultanément réalisées :

- les deux prédicteurs sont « suffisamment décorrélés »: $\rho_{12} < 0$
- la variable à prédire y est « suffisamment corrélée » avec la première composante principale (de valeur propre positive) c'est-à-dire avec coefficient de corrélation supérieur à $\sqrt{2}/2$

Cette propriété sera démontrée plus loin avec une démonstration géométrique.

2.2.3 Importance *beta square*

Cette méthode consiste simplement à affecter à chaque prédicteur une valeur d'importance relative qui correspond au carré du coefficient standardisé de l'OLS.

Définition : Importance *betasq*

L'importance betasq est définie comme :

$$I_{betasq}(j) = b_j^2 \left(\frac{s_j^2}{s_y^2} \right) = \beta_j^2 \qquad (2.2.3.1.)$$

Cette décomposition a été fortement critiquée par Johnson et al. (2004) et elle a été largement abandonnée dans la mesure où elle ne conduit pas non plus à une décomposition complète du R² puisque la somme des carrés des β standardisés n'est pas égale au carré de la somme expliquée en cas de corrélation entre les prédicteurs.

Rappelons néanmoins que la méthode de référence telle que présentée par IBM SPSS (cf. chapitre 2 et bibliographie) recommande justement d'utiliser les coefficients standardisés. Outre le fait que les carrés des coefficients standardisés ne forment pas le R^2 quand ils sont additionnés (sauf dans le cas d'indépendance des

prédicteurs), les rapports relatifs entre les β_j et leurs carrés respectifs sont naturellement très différents. Nous reviendrons sur ce point à propos de l'analyse de la décomposition par la valeur de Shapley.

Notons cependant que si les prédicteurs sont tous corrélés positivement avec la variable à prédire, ce qui est un cas assez plausible dans le cas des études de marchés, comme la fonction $x \rightarrow x^2$ est monotone et croissante sur $\mathbf{R}^+$ le rangement d'importance de la fonction d'importante *betasq* sera identique à celui de l'OLS.

Cette propriété sera bien sûr conservée si la condition de signe est respectée et que les prédicteurs sont mutuellement décorrélés.

Donc la décomposition « last » est conforme au sens évoqué plus haut pour des familles de variables où les X sont mutuellement décorrélés.

A titre d'illustration, voici les carrés des β_j^2 standardisés pour le jeu de données *swiss* :

Agriculture	Examination	Education	Catholic	Infant.Mortality
0,09792	0,02715	0,44944	0,12083	0,06307

Tableau 2.2.3.1 Importances *betasq*. Données *swiss*.

Dans sa synthèse Grömping (2006) après avoir présenté les trois méthodes qui précèdent, décrit la décomposition de Pratt, puis introduit deux autres décompositions de la variance, *lmg* que nous appellerons dans cette recherche *lmg-Shapley* et *pmvd*. Grömping souligne le caractère « computer-intensive » de ces deux dernières approches. Nous allons consacrer un chapitre particulier à la décomposition de la variance vu son rôle accru depuis une quinzaine d'années dans les études de marchés, souligné encore récemment par Conklin et Lipovetsky (2013).

Chapitre 3 : Décomposition de la variance

3.1 Décomposition de la variance

La décomposition de la variance a été proposée par plusieurs praticiens d'études de marchés. Dans d'autres secteurs d'activité la décomposition de la variance a également été utilisée pour quantifier la relative importance des prédicteurs : sciences humaines, mathématiques financières. Grömping (2007) a présenté le cadre général de la décomposition de la variance dans le cas du modèle linéaire.

Nous allons la décrire ci-après dans le cadre de l'utilisation de la régression linéaire sur un ensemble d'observations. Il est utile de présenter une formule qui joue un rôle clé dans l'interprétation de ces approches. En reprenant les notations introduites en tête du rapport, la variance expliquée $V(\mathbf{y}^*)$ peut s'écrire, en notant $r(\mathbf{x_i}, \mathbf{x_j})$ le coefficient de Bravais-Pierson entre $\mathbf{x_i}$ et $\mathbf{x_j}$ et s_i l'écart-type de $\mathbf{x_i}$:

$$V(\mathbf{y}^*) = V(\sum_{i=1}^{i=p} b_i \mathbf{x_i}) = \sum_{i=1}^{i=p} b_i^2 V(\mathbf{x_i}) + 2\sum_{i<j} r(\mathbf{x_i}, \mathbf{x_j}) b_i b_j s_i s_j \quad (3.1.1)$$

En notant $R^2(P)$ le coefficient de détermination du modéle linéaire obtenu en prenant comme prédicteurs les variables de P, une fonction d'importance sera une décomposition du R^2 si :

$$\sum_{j=1}^{j=p} I(j) = R^2(P) \quad (3.1.2)$$

Plusieurs approches consistent à partir des termes de l'équation (3.1.1) et à en déduire des quantificatinos de l'importance.

Dans ce chapitre et afin de simplifier les notations nous présenterons les méthodes des formules de décomposition du R^2 en utilisant les coefficients de régression standardisés notés β_i pour chaque prédicteur i.

Afin de comparer plus précisément les différentes méthodes nous allons définir le vecteur de rangement d'importance.

Définition : Rangement d'importance

Soit une fonction d'importance et soit un échantillon d'observations de p prédicteurs et d'une variable à prédire. Chaque prédicteur reçoit une valeur d'importance par la fonction d'importance I. Soit I(P) l'ensemble de ces valeurs. I(P) est un ensemble fini d'au plus p valeurs distinctes de $\mathbf{R}$. *A chaque variable de j de 1 à p nous pouvons associer un rang r(j) par ordre décroissant avec la convention que si deux variables ou plus ont la même importance elles auront le même rang et le nombre de rangs sera alors strictement inférieur à p. Le Rangement d'Importance est le vecteur* $(r_1,...r_p)'$.

Nous noterons $RGI(\mathbf{y},\mathbf{X},\mathbf{I})$ le rangement d'importance observé sur l'échantillon et associé aux variables $\mathbf{y},\mathbf{X}$ et à la fonction d'importance I. Dans un souci de simplicité nous noterons de la même façon le rangement des importances estimées sur un jeu de données.

Cette définition permet de formaliser la comparaison des rangs entre deux méthodes appliquées à un même ensemble de prédicteurs où les rangs correspondant aux estimateurs d'importance sur un jeu de données particulier.

3.1.1 Décomposition de Pratt

Hoffmann (1960) et Pratt (1987) ont proposé de multiplier le coefficient standardisé de l'OLS par la corrélation marginale.

Définition : Importance *pratt* :

$$\mathrm{I}_{pratt}(j) = r(\mathbf{y},\mathbf{x_j})\beta_j \tag{3.1.1.1}$$

Comme montré plus haut avec l'équation (2.1) l'importance *pratt* est une décomposition de la variance et les valeurs $\mathrm{I}_{pratt}(j)$ respectent l'équation (3.1.3). Les termes $\mathrm{I}_{pratt}(j)$ sont appelés SCCP : (*Standardised Coefficient Correlation Products).*

Lipovetsky et al. (2001) ont proposé de prendre en compte les impacts potentiels indirects des autres prédicteurs pour calculer des « *net effects* » définis comme ci-après :

$$NE(i) = \beta_i{}^2 + \sum_{\mathbf{i}\neq\mathbf{j}} \beta_i \beta_j r(\mathbf{x_i},\mathbf{x_j}) \tag{3.1.1.2}$$

Cette formule est en fait la traduction d'un modèle qui suppose que l'effet d'un prédicteur (ici le prédicteur i) influerait sur la variable Y non seulement directement (c'est le terme β_i^2) mais aussi via sa corrélation avec les autres prédicteurs (impacts « indirects ») qui sont les termes :

$$\sum_{\mathbf{i}\neq\mathbf{j}} \beta_i \beta_j r(\mathbf{x_i},\mathbf{x_j})$$

La partition de la variance expliquée par les Net Effects consiste à ajouter ces deux termes pour calculer la part de variance allouée au prédicteur i en additionnant :

- la part de variance β_i^2
- et la moitié de la variance « conjointe » entre les prédicteurs i et j c'est-à-dire la moitié de $2r(\mathbf{x_i}, \mathbf{x_j})\beta_i\beta_j$ dans le développement algébrique de la variance expliquée.

Nous avons ainsi identifié un lien entre les Net Effects et la décomposition *pratt.*

Résultat : *Pour chaque prédicteur les Net Effects sont identiques aux termes d'importance de la décomposition de Pratt.*

En effet :

$$\mathrm{I}_{pratt}(\mathrm{j}) = r(\mathbf{y}, \mathbf{x_j})\beta_j = NE(j) \qquad (3.1.1.3)$$

Voici les résultats toujours avec les données *swiss* :

Agriculture	Examination	Education	Catholic	Infant.Mortality
-0,110486	0,1064274	0,4450046	0,1611768	0,1046122

Tableau 3.1.1.1 Importances *pratt.* Données *swiss*.

Cette métrique a été critiquée dans la mesure où elle peut conduire à des contributions négatives. Lipovetsky et al. (2001) considèrent que l'apparition d'un terme négatif dans la décomposition de Pratt serait le fait d'imprécisions statistiques. Grömping et Landau (2009) relèvent qu'il peut exister des configurations structurelles où un β standardisé peut être négatif alors que la corrélation marginale sera positive. Nous partageons ici le point de vue de Grömping et Landau, car il peut être parfaitement possible au plan structurel et donc légitime au plan du modèle d'analyse, de rencontrer pour un prédicteur donné un coefficient de régression multiple et un coefficient de corrélation avec la variable à prédire de signes différents. Même si cette situation peut paraître contre-intuitive pour les résultats d'études de marchés, elle ne saurait être exclue *a priori*.

Pour s'en convaincre nous proposons ci-après une interprétation géométrique dans le cas de la régression linéaire c'est-à-dire en considérant un jeu d'observations avec les notations présentées précédemment. Considérons le cas de deux prédicteurs moyennement corrélés, il suffit que la projection de la variable à prédire se situe à l'extérieur de l'angle aigu formé par les deux prédicteurs mais avec une corrélation positive avec les variables $\mathbf{x_1}$ et $\mathbf{x_2}$ et nous pouvons avoir :

- les corrélations entre la variable à prédire et chacun des prédicteurs qui sont positives

- mais un des prédicteurs (celui qui aura avec la projection de la variable à prédire sur le plan $(\mathbf{x_1}, \mathbf{x_2})$ la corrélation la plus faible) se verra attribuer un coefficient négatif.

Cette situation est voisine de la configuration de suppression telle que définie par Cohen et al. (2003) : ces auteurs expliquent qu'il y a situation de suppression dans les cas suivants : (en reprenant les notations de Cohen et al. c.à.d. $r_{12} = r(\mathbf{x_1}, \mathbf{x_2})$, $r_{yj} = r(\mathbf{y}, \mathbf{x_j})$) :

- $r_{y1} \leq r_{y2} r_{12}$
- $r_{y2} \leq r_{y1} r_{12}$
- $r_{12} \leq 0$

Comme : $\beta_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$ et $\beta_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$ et aussi

$r_{y1} = \beta_1 + r_{12}\beta_2$ et $r_{y2} = r_{12}\beta_1 + \beta_2$

dire comme les auteurs que le terme d'« effet direct » doit en principe inférieur au terme de « zero order correlation » peut s'écrire :

$$\beta_1 \leq r_{y1}$$

Ou encore :

$$\beta_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \leq r_{y1}$$

Soit :

$$r_{12}^2 r_{y1} \leq r_{12} r_{y2}$$

Dans le cas où $r_{12} \geq 0$, la condition $r_{12} r_{y1} \leq r_{y2}$ correspond au cas où $\beta_2 \geq 0$. La situation de suppression du prédicteur 1 sera ainsi rencontrée dans un cas comme l'exemple précédent ou l'un seulement des β (ici pour le prédicteur 1) est négatif.

Dans les études sociales et marketing il est souvent tentant de rejeter du modèle des variables à coefficients négatifs ou par exemple d'ajuster les coefficients même significativement négatifs à une petite valeur positive et dans les cas les plus rigoureux d'application à vérifier par un test la non significativité de la perte de R^2.

Voici, en complément de l'approche géométrique ci-dessus, une illustration en appliquant la régression linéaire à un jeu de données. Soient deux variables aléatoires sur $\mathbf{R}$ indépendantes et d'espérance nulle représentées par deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ de $\mathbf{R}^n$ avec $E(\mathbf{u}) = E(\mathbf{v}) = 0$ et $r(\mathbf{u}, \mathbf{v}) = 0$.

Il est possible de construire deux prédicteurs $\mathbf{x_1}$ et $\mathbf{x_2}$ comme des combinaisons linéaires de $\mathbf{u}$ et $\mathbf{v}$, et une variable à prédire $\mathbf{y}$ où les corrélations entre $\mathbf{y}$, $\mathbf{x_1}$ et $\mathbf{x_2}$ sont toutes deux à deux positives mais où le coefficient b_i dans la régression linéaire de $\mathbf{y}$ sur $(\mathbf{x_1}, \mathbf{x_2})$ d'un des prédicteurs est négatif :

Posons :

$$\mathbf{x_1} = \cos(\varphi)\mathbf{u} - \sin(\varphi)\mathbf{v}$$
$$\mathbf{x_2} = \cos(\varphi)\mathbf{u} + \sin(\varphi)\mathbf{v}$$
$$\mathbf{y} = \cos(\psi)\mathbf{u} + \sin(\psi)\mathbf{v}$$

En choisissant par exemple $\varphi = 21$ degrés et $\psi =$ 40 degrés, il vient :

$$r(\mathbf{y}, \mathbf{x_1}) = 0,48$$
$$r(\mathbf{y}, \mathbf{x_2}) = 0,95$$
$$r(\mathbf{x_1}, \mathbf{x_2}) = 0,48$$
$$\beta_1 = -0,49$$
$$\beta_2 = 1,30$$

Donc pour le prédicteur $\mathbf{x_1}$ le coefficient de corrélation bivariée avec $\mathbf{y}$ est de signe opposé au coefficient de la régression multiple de $\mathbf{y}$ sur $(\mathbf{x_1}, \mathbf{x_2})$.

En conséquence il serait erroné avec ces trois variables de mettre l'apparition sur certains échantillons de coefficients négatifs sur le compte d'imprécisions statistiques. Les estimateurs de β_i étant asymptotiquement convergents il existe des échantillons tels que les estimateurs du coefficient de corrélation et du coefficient standardisé soient de signes contraires, et significativement non nuls.

Dans ces conditions l'apparition de termes négatifs dans la décomposition d'Hoffmann peut donc résulter de la structure même des données ou de la multiplicité des prédicteurs, dans la mesure où le prédicteur recevant une valeur de Hoffmann négative peut avoir un coefficient de corrélation bivariée positif lorsqu'il est utilisé seul dans le modèle mais un coefficient négatif dans la régression multiple avec d'autre prédicteurs. Pour contourner l'apparition de termes négatifs quelques variantes ont été proposées (cf. Menard (2007) dans une réponse à l'article de Grömping (2007). Ces variantes consistent à prendre comme valeur d'importance pour chaque prédicteur la valeurs absolue du produit du coefficient standardisé de la régression multiple par le coefficient de la corrélation avec la variable à prédire ; c'est l' ASCCP (Absolute Standardized Coefficients Correlation Products) ou aussi afin que la somme de ces termes soit bien égale au R^2, à effectuer une normalisation des ASCCP : Menard parle alors de NSCCP pour Normalized Stantardised Coefficients Correlations Products.

En raison des limitations et du caractère artificiel de ces dernières options, les chercheurs travaillant sur la décomposition de la variance se sont plutôt intéressés à des techniques de décomposition plus complexes en termes de calculs en incorporant des sélections ordonnées de variables avec des formules consistant à « moyenner » (« averaging ») les écarts de R^2 sur les différentes séquences de classement (« ordering ») des prédicteurs. Ces méthodes ont été rendues possibles par la capacité croissante des ordinateurs et la diffusion de packages R, techniques qui n'étaient pas accessibles quand Hoffmann s'est intéressé au sujet.

Deux méthodes de décomposition fondées sur des calculs d'accroissements marginaux du R^2 sont principalement référencées : *lmg-Shapley Value* et *pmvd* qui vont être maintenant présentées.

3.1.2 lmg ou Shapley Value

Cette méthode a été introduite initialement par Lindeman, Merenda et Gold (1980), d'où l'acronyme *lmg*. Le principe est de calculer une moyenne sur tous les ordonnancements possibles de prédicteurs l'accroissement R^2 quand ce prédicteur est ajouté aux prédicteurs situé après lui dans chacun des ordonnancements. Vu sa diffusion depuis plusieurs années dans les études de marchés nous allons y consacrer une présentation plus détaillée.

La formule de calcul de l'importance de de chaque prédicteur selon l'approche *lmg-Shapley* est donnée ci-après : pour le prédicteur j, $I_{lmg}(j)$ est définie comme la moyenne des accroissements du R^2 sur l'ensemble des sous-modèles contenant le prédicteur j par rapport au même sous-modèle ne contenant pas le prédicteur j et ce en considérant que les variables sont incluses successivement selon tous les ordres possibles des prédicteurs donc avec p ! ordonnancements possibles. Le cas de deux et trois prédicteurs illustre bien la méthode :

Cas de deux prédicteurs :

Il y a deux ordonnancements : 1,2 et 2,1.

L'importance pour le prédicteur 1 sera :

$$lmg(1) = \frac{1}{2}\{(R^2(1,2) - R^2(2)) + (R^2(1) - R^2(\varnothing))\}$$

L'importance pour le prédicteur 2 sera :

$$lmg(2) = \frac{1}{2}\{(R^2(1,2) - R^2(1)) + (R^2(2) - R^2(\varnothing))\}$$

Comme $R^2(\varnothing) = 0$ on a bien $lmg(1) + lmg(2) = R^2(1,2)$

Cas de 3 prédicteurs :

il y a 6 ordonnancements possibles et nous pouvons par exemple calculer *lmg* pour la variable 1 :

$$lmg(1) = \frac{1}{6}\{(R^2(1,2,3) - R^2(2,3)) + (R^2(1,3,2) - R^2(3,2)) + (R^2(1,2) - R^2(2)) + (R^2(1,3) - R^2(3)) + (R^2(1) - R^2(\varnothing))$$

De façon générale considérons maintenant le cas d'un prédicteur j parmi un ensemble P de p prédicteurs et soit $K \subset P$, et notons k le nombre d'éléments de K , $k = card(K)$.

Nous pouvons remarquer que pour un ordonnancement donné l'accroissement du R^2 quand le prédicteur j est ajouté au modèle ne dépend pas de l'ordre des prédicteurs qui le précèdent, ni de ceux qui le suivent. Si K désigne dans un ordonnancement donné l'ensemble des prédicteurs qui précèdent j parmi les $p!$ ordonnancements possibles $k!(p-k-1)!$ donnent donc le même accroissement de R^2

I_{lmg} peut être formulée de la façon suivante :

Définition : Importance *lmg*

$$I_{lmg}(j) = \sum_{K \subset P/\{j\}} \frac{k!(p-k-1)!}{p!}\left[R^2(K \cup \{j\}) - R^2(K)\right] \qquad (3.1.2.1)$$

Il résulte de cette définition que I_{lmg} est une décomposition de variance. (Cf. Grömping (2007).

$$\sum_j I_{lmg}(j) = R^2$$

L'approche *lmg* s'avère en fait identique dans ses résultats avec ceux obtenus en utilisant une approche dérivée de la théorie des jeux, la Shapley Value. Nous allons d'abord rappeler la définition et les propriétés de la Shapley Value en théorie des jeux et montrerons le lien avec *lmg* et l'analyse des leviers.

En théorie des jeux un jeu coopératif est un jeu tel que les joueurs ont la possibilité de se concerter et de s'engager à coopérer avant de définir la stratégie à adopter.

Considérons un jeu coopératif entre un ensemble P de p joueurs. Soit une fonction de gain v qui associe à chaque coalition de joueurs (sous -ensemble de P) une valeur de gain dans $\mathbf{R}$ $v(P)$.

Définition : Jeu Coopératif

Un jeu coopératif est un couple (P, v) avec

- *P est un ensemble fini de joueurs*
- *v est une fonction de* $2^P \to \mathbf{R}$ *telle que v(Ø)=0 et si A est inclus dans B alors v(A) ≤v(B)*

Définition : Fonction de Valeur

Une fonction de valeur (value fonction) est une fonction qui associe à un jeu coopératif

(P , v) un vecteur dans $\mathbb{R}^P$ $\left(f(1), \ldots . f(p) \right)$.

Le lien entre *lmg* et la théorie des jeux vient de ce que :

- chaque prédicteur j est assimilé à un joueur dans un jeu coopératif
- La fonction de gain du jeu coopératif est définie pour chaque coalition (sous-ensemble) $K \subseteq P$ par $v(K) = R^2(K)$, la variance expliquée en utilisant les prédicteurs inclus dans K .

En voici une formalisation issue de la théorie des jeux.

Définition : Contribution marginale :

Soit un jeu coopératif entre p joueurs en une fonction de gain v. Soient S une coalition et un joueur $j \notin S$ *. La contribution marginale du joueur* j *à la coalition S est :*

$$\Delta_j(S) = v(S \cup \{j\}) - v(S)$$

Définition : Contribution marginale ordonnée

Soit π une permutation de $P = \{1, \ldots p\}$ *et un joueur* $j \in P$ *et* $\pi(l)$ *l'image du prédicteur* l *par la permutation* π *. Soit* p_π^j *le sous ensemble des prédicteurs tels que* $\pi(l) > \pi(j)$. *La contribution ordonnée de j dans la permutation π est la contribution marginale de j à la coalition* p_π^j.

$$\Delta_j(\pi) = v(p_\pi^j \bigcup \{j\}) - v(p_\pi^j)$$

Définition : Shapley Value

Soit Π l'ensemble des permutations de {1,2,…p}, l'allocation Shapley Value est l'application qui à chaque joueur j associe $\varphi(j)$ *tel que :*

$$\varphi(j) = (1/p!)\sum_{\pi\in\Pi}\Delta_j(\pi)$$

Elle associe donc à chaque joueur la moyenne sur toutes les contributions marginales ordonnées possibles. Nous admettrons ici le résultat suivant (Shapley, (1953)) : la Shapley Value est la seule fonction de valeur qui respecte simultanément les axiomes d'efficacité, de symétrie, nullité et additivité tels que détaillés ci-après :

Efficacité

$$\sum_{i=1}^{p} f(i) = v(P)$$

Symétrie

Deux joueurs sont symétriques dans un jeu (P, v) si leurs contributions marginales sont identiques par rapport à toute coalition. La propriété de symétrie est :
i et j symétriques $\Rightarrow$ f(i)= f (j)

Nullité

Un joueur sera dit joueur nul si sa contribution marginale à toute coalition du jeu est nulle. La propriété de nullité s'écrit : j joueur nul entraîne f(j)=0

Additivité

Soient deux jeux (P, v) et (P, w), définissons le jeu (P, v+w) par v+w(K)=v(K) +w(K)

La propriété d'additivité est f(v+w)=f(v)+f(w).

La formule explicite de I_{lmg} nous montre que *lmg* est exactement identique à Shapley Value avec comme jeu (P, R^2) c'est-à-dire $v(K) = R^2(K)$ pour tout $K \subset P$.

Le fait que la décomposition de la variance via la Shapley Value ait un lien avec la théorie des jeux a pu contribuer à lui donner une crédibilité chez certains auteurs (Pokryshevskaya et Antipov (2013)) qui expliquent que l'importance des attributs a été réalisée en utilisant la décomposition du R^2 par la Shapley Value *« qui a été introduite récemment en économétrie à partir de la théorie des jeux. De ce fait l'approche par la Shapley Value est justifiée au plan théorique »* (page 6).

Notons cependant que nous parlons bien là d'un mécanisme d'allocation d'un gain total constant (le R^2 obtenu en prenant toutes les variables) et que ceci est un cas bien particulier au sens ou l'ajout d'une nouvelle variable très corrélée avec les autres prédicteurs n'apporte aucun gain alors que dans les modèles de jeux économiques de production par exemple il y a amélioration du gain total. A l'extrême si une variable est introduite deux fois dans le jeu, le R^2 avec toutes les variables n'est pas changé : il n'y a pas de gain supplémentaire. Nous savons aussi que ces deux variables identiques recevront la même valeur de Shapley (axiome de symétrie ou calcul direct).

Prenons à titre de comparaison un exemple d'utilisation de la Shapley Value tiré des travaux d'Alexis Eidelman (2012) avec là aussi des joueurs (des ouvriers au lieu de considérer des variables prédictives comme joueurs).

Dans une entreprise avec n-1 ouvriers que produisent « 1 » chacun quand il n'y a pas de manager, et « 2 » quand il y a un manager : p ouvriers seuls produisent p et p ouvriers avec le manager produisent 2p et ce quel que soit p entre 1 et n-1. Aussi par hypothèse le manager seul ne produit rien. Dans cet exemple la Shapley Value du manager sera (n-1)/2, chaque ouvrier se voyant attribuer 3/2, donc au total $3/2\,(n-1)$ pour l'ensemble des ouvriers et bien sûr comme attendu le gain total est bien 2(n-1).

Nous voyons ici une différence essentielle car l'addition d'un nouvel ouvrier ne diminue pas la Shapley Value allouée qui reste ½ tandis que dans l'application à la décomposition de la variance la duplication d'une même variable vient automatiquement diminuer la variance allouée à cette variable initialement seule.

Ceci signifie simplement que le « jeu » représenté par l'allocation de variance entre variables ne s'enrichit pas si des variables sont introduites plusieurs fois, contrairement à des applications du type microéconomique. En ce sens l'utilisation de la puissante référence qu'est la théorie des jeux nécessite une précaution d'interprétation.

Cependant la Shapley Value a rapidement gagné en notoriété dans l'industrie des études de marchés avant que Lipovetsky et al. (2001) n'y apportent une étape additionnelle aboutissant à recalculer des β après une décomposition intermédiaire de la variance.

Avec la méthode *lmg-Shapley* nous avons cette fois une méthode qui décompose bien le R^2 et ne fournit pas de termes négatifs ces deux premiers critères étant cités par différents auteurs mentionnés ci-après comme souhaitables dans la décomposition de la variance.

Deux autres critères ont été également mis en avant par ces auteurs (Johnson and Lebreton (2004), Feldman (2005),

- **Exclusion** : un prédicteur avec un $\beta = 0$ devrait recevoir une allocation de variance nulle par la méthode de décomposition considérée.

- **Inclusion** : un prédicteur avec un $\beta \neq 0$ devrait recevoir une allocation non nulle.

Nous appellerons dans la suite les importances allouées $SV(i)$ pour le prédicteur i , SV pour Shapley Value, ou indifféremment $lmg(i)$.

Notons que la méthode *lmg-Shapley* ne respecte pas le critère d'exclusion. Ceci peut être démontré simplement dans le cas de deux prédicteurs (Grömping (2007)) :

$$lmg(1) = (\beta_1^2) + (\beta_1\beta_2 r(\mathbf{x_1}, \mathbf{x_2}) + 0,5(\beta_2^2 - \beta_1^2) r^2(\mathbf{x_1}, \mathbf{x_2}) \qquad (3.1.2.2)$$

Ecrivons d'une autre manière le résultat présenté par Grömping ;

$$SV(1) = \frac{R^2 + (\beta_1^2 - \beta_2^2)(1 - r^2(\mathbf{x_1}, \mathbf{x_2}))}{2}$$

Avec le cas $\beta_1 = 0$ *et* β_2 donné différent de 0 il apparaît néanmoins que $SV(1)$ est différent de zéro. Le critère d'exclusion n'est donc effectivement pas respecté.

Résultat : *La décomposition par la Shapley Value ne conserve pas le rangement d'importance par rapport au classement établi selon les coefficients standardisés du modèle linéaire simple.*

Ce résultat découle du terme suivant dans le cas de deux prédicteurs :

$$SV(1) = \frac{R^2 + (\beta_1^2 - \beta_2^2)(1 - r^2(\mathbf{x_1}, \mathbf{x_2}))}{2}$$

Il s'ensuit que :

$$SV(1) - SV(2) = (\beta_1^2 - \beta_2^2)(1 - r^2(\mathbf{x_1}, \mathbf{x_2}))$$

Il est donc possible de générer une configuration avec trois prédicteurs avec SV(1), SV(2) et SV(3) dans un ordre différent des β standardisés.

Voici un exemple théorique qui permet de comprendre le mécanisme de changement d'ordre :

Considérons trois variables centrées réduites (V_1, V_2, V_3) avec :

- V_3 décorrélé de V_1 et V_2 : $\rho_{31} = \rho_{32} = 0$
- et $\rho_{12} = 0,6$.

Soit également une variable à prédire y définie par $\beta_1 = 1,1$; $\beta_2 = 0,7$; $\beta_3 = 0,9$

L'ordre entre les betas et les Shapley Values est diffèrent à cause de la réallocation ente V1 and V2 par la Shapley Value comme montré ci-dessous. Ceci signifie que l'utilisation directe la Shapley Value peut changer l'ordre par rapport à l'application de l'OLS.

	Beta 1	Beta2	Beta3
Valeur	1.1	0.7	0.9
Rang:	1	3	2

Tableau 3.1.2.1: OLS

	SV1	SV2	SV3
Valeur	1.54	1.08	0.81
Rang :	1	2	3

Tableau 3.1.2.2: *lmg-Shapley Value*

Résultat : *Soient deux prédicteurs i et j tels que* $\beta i \neq 0$ *et* $\beta j \neq 0$ *.Dans le cas où les prédicteurs sont mutuellement décorrélés les identités suivantes sont respectées : pour i et j*

$$\forall i, j : (SV(i) / SV(j)) = (\beta_s^i / \beta_s^j)^2$$

Ceci découle directement de l'équation (3.1.2.2) et de la définition des coefficients standardisés.

Par continuité dans le cas de non orthogonalité modérée des prédicteurs, les ratios d'importance issus de la Shapley Value sont donc susceptibles de différer fortement des ratios d'importance issus des β standardisés.

Bien que les publications techniques aient pris la précaution d'indiquer que les SV ne sauraient être utilisés directement pour la prédiction (ceci est aussi clairement rappelé par Grömping) et s'efforcent de bien distinguer la simulation (ou prédiction) de la décomposition de la variance il est tentant pour les praticiens et naturel pour des utilisateurs non spécialistes de présenter et d'utiliser les résultats de cette façon , et si un prédicteur reçoit une SV 4 fois plus élevée qu'un autre de simuler implicitement une régression avec ces coefficients.

Comme le montre le résultat ci-dessus le rapport des SV peut en réalité être très différent du rapport des β.

Sur les deux jeux expérimentaux que nous avons utilisés, si les valeurs de Shapley sont utilisées comme coefficients dans le modèle linéaire, nous avons observé une diminution significative du R^2 confirmée par un F test (cf. Ci-après). De façon générale les coefficients générés par la décomposition par la Shapley Value ne permettent pas de réaliser efficacement une simulation.

Comme évoqué précédemment la Shapley Value est une moyenne de différences entre termes du type suivant :

- $R^2(K \cup \{j\})$ avec le modèle incluant les prédicteurs de K et le prédicteur j.
- $R^2(K)$ avec le modèle incluant les prédicteurs de K sans le prédicteur j.

Cette différence est donc le « last » de X_j dans le sous-modèle de la variable y avec les variables de l'ensemble $K \cup \{j\}$.

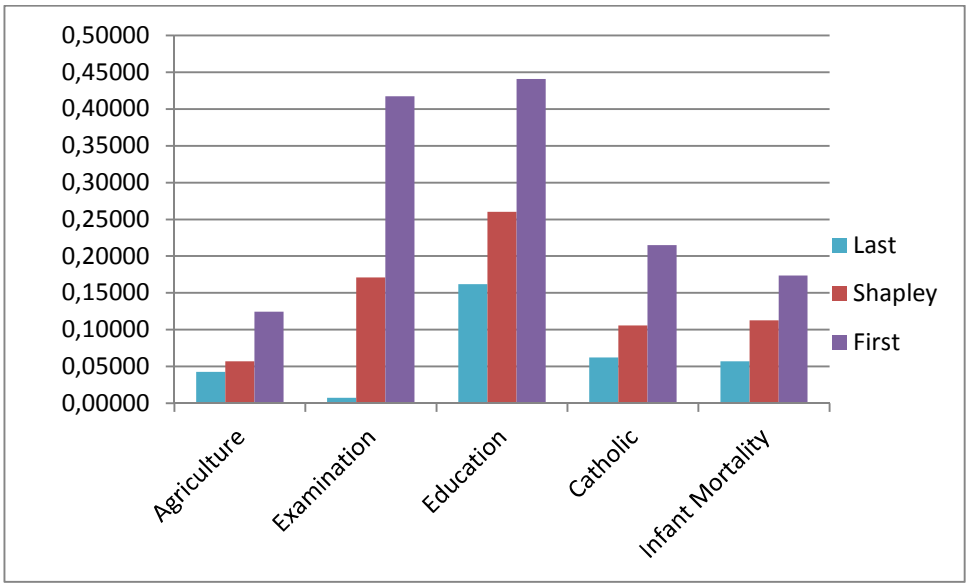
Une autre manière d'écrire le terme $R^2(K \cup \{j\}) - R^2(K)$ est de noter qu'il s'agit du carré de la corrélation semi partielle de j pour la variable à prédire y dans le modèle linéaire avec les variables de l'ensemble $K \cup \{j\}$

Donc :

$$R^2(K \cup \{j\}) - R^2(K) = \beta_j^2 Tol_j$$

avec $\text{Tol}_j = 1 - R^2_{XjK}$, Tolérance de X_j vis-à-vis des variables de $K \cup \{X_j\}$

A titre d'illustration le tableau suivant montre la hiérarchie entre *last, lmg-Shapley Value* et *first* dans le cas des données « *swiss* » :



Graphique 3.1.2.1 : Comparaison entre *last*, *lmg –Shapley,* et *first.* Données *swiss.*

Voici aussi comme pour les autres méthodes les résultats de la décomposition lmg-Shapley dans le cas des données *swiss* :

Agriculture	Examination	Education	Catholic	Infant.Mortality
0,05709122	0,17117303	0,26013468	0,10557015	0,11276592

Tableau 3.1.2.3. Importances *lmg-Shapley*. Données *swiss*.

Une approche supplémentaire utilisant les valeurs de la décomposition de Shapley est la « Dominance Analysis ». Cette approche consiste à calculer pour chaque prédicteur une deuxième valeur qui est la différence *first -lmg*. La Dominance Analysis ne se limite pas à calculer ces deux valeurs, mais aussi à comparer toutes les paires de prédicteurs. Par définition le prédicteur A aura une « General dominance » sur le prédicteur B si lmg(A) est supérieur à lmg(B). A aura une « Complete dominance » sur B si cette inégalité est vérifiée quelles que soient les autres variables dans les sous-modèles. Grömping (2015) ne recommande pas l'utilisation de la « Complete dominance », mais considère que la « General Dominance » est un complément utile à la décomposition *lmg-Shapley*. Cette approche n'est pas à notre connaissance utilisée dans les études de marché. Elle présente naturellement un lien avec la comparaison des *firsts* et des valeurs de *lmg-Shapley*.

3.1.3 Décomposition *pmvd*

La décomposition *pmvd* (Proportional Marginal Variance Decomposition) a été introduite par Feldman (1999) et vise à respecter le critère d'inclusion mentionné précédemment au 3.1.2. C'est-à-dire que si un prédicteur j a un β_j nul sa part de variance allouée sera bel et bien nulle, contrairement par exemple au cas de *lmg-Shapley* comme montré précédemment dans le cas de deux prédicteurs.

Le calcul de pmvd(j) pour un prédicteur donné ressemble au calcul précédent pour *lmg-Shapley*.

Nous allons le détailler ci-après de façon harmonisée en prenant le formalisme de la théorie des jeux.

Assimilons les p prédicteurs à p joueurs et définissons le gain d'une coalition de joueurs comme le R^2 obtenu par le modèle incluant les prédicteurs dans la coalition et $\Delta_j(r)$ comme l'accroissement marginal (cf. chapitre 2.3.2) du prédicteur j par rapport aux prédicteurs qui le précèdent dans la permutation r (cf. Grömping (2007) et Feldman (1999 et 2005)).

$$pmvd(j) = \frac{1}{p!} p(r) \sum_r \Delta_j(r)$$

$$lmg(j) = \frac{1}{p!} \sum_r \Delta_j(r)$$

Les deux formules précédentes sont proches et sont calculées de façon formellement très voisines mais pmvd(j) inclut un terme spécifique $p(r)$. Si nous posons pour une permutation $r = r_1, ...r_i, ...r_P$ des p prédicteurs les définitions suivantes (cf. Grömping (2007)):

$$e\,\mathrm{var}(y) = \mathrm{var}(y) - \mathrm{var}(y \,|\, Xj, j \in S)$$

$$L(r) = \prod_{i=1}^{p-1} (e\,\mathrm{var}(\{1,..., p\} - e\,\mathrm{var}(\{r_1,..., r_i\}))^{-1}$$

$$p(r) = \frac{L(r)}{\sum_r L(r)}$$

Si certains coefficients β sont nuls, leurs pmvd seront nuls et les résultats issus des données seront identiques aux modèles ou ces variables sont omises (cf. Grömping (2007)).

Dans le cas de deux prédicteurs le calcul donne le résultat suivant :

$$pmvd(1) = \beta_1^2 + \frac{\beta_1^2}{\beta_1^2 + \beta_2^2} 2\beta_1 \beta_2 r^2(\mathbf{x_1}, \mathbf{x_2})$$

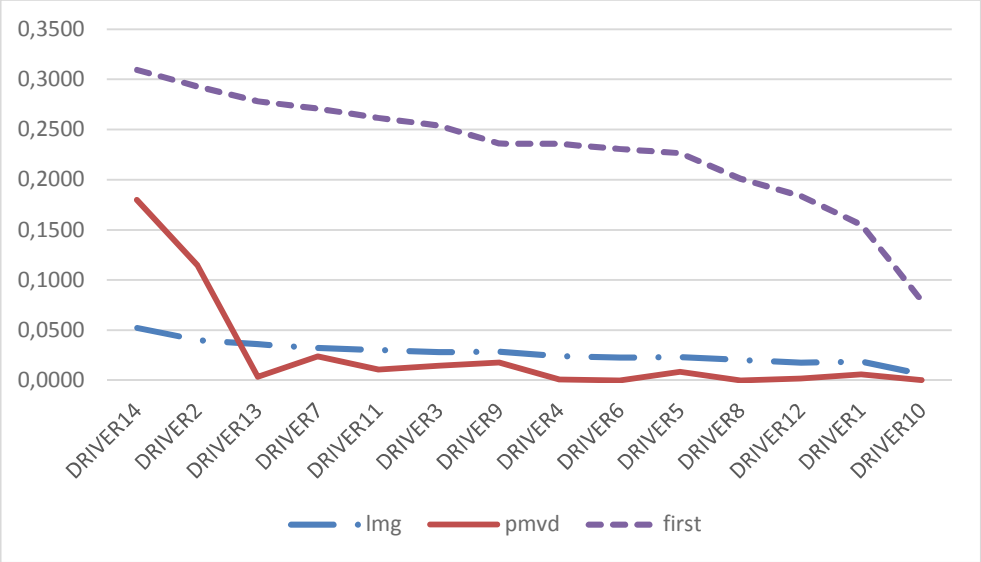
$$pmvd(2) = \beta_2^2 + \frac{\beta_2^2}{\beta_1^2 + \beta_2^2} 2\beta_1 \beta_2 r^2(\mathbf{x_1}, \mathbf{x_2})$$

Comme prévu cette fois si β_j est nul alors $pmvd(j)$ l'est aussi, donc la propriété d'exclusion est vérifiée. Voici les résultats tels que présentés par Grömping (2006) pour les données *swiss* avec le package R `relaimpo`. (*)

Agriculture	Examination	Education	Catholic	Infant.Mortality
0,04478517	0,04446868	0,37981877	0,13433174	0,10333064

Tableau 3.1.3.1 Importances *pmvd.* Données *swiss*.

Le graphique ci-dessous présente dans ce cas de *UK Data* une comparaison des méthodes *lmg-Shapley, pmvd* et *first.*



Graphique 3.1.3.1 Comparaison *lmg-Shapley, pmvd* et *first*. Données *UK Data.* Importances non normalisées.

La décomposition *pmvd* nécessite des calculs plus lourds que *lmg* et par ailleurs les valeurs *pmvd* présentent une plus grande variabilité que celles obtenues par *lmg* (Grömping 2007).

(*) : *A noter que dans le package* `relaimpo` *le code de calcul de pmvd n'est accessible que pour les utilisateurs en dehors des Etats-Unis en raison d'une protection de la propriété intellectuelle dans ce pays.*

3.1.4 Décomposition d'Owen

La décomposition d'Owen est une généralisation de la décomposition de Shapley qui prend en compte des associations entre joueurs. La Shapley Value a été introduite par Shapley en 1953. La Owen value par Owen en 1977, également dans le cadre de la théorie des jeux, la Owen value prenant en compte des coalitions entre joueurs. L'application à la décomposition du R^2 inspirée par la Shapley Value a été étendue via le recours à la Owen Value, avec l'idée d'effectuer des regroupements *a priori* entre variables. Frank Huettner et Marco Sunder, de l'université de Leipzig ont publié en 2011 et 2012 des articles présentant cette méthode et développé un module `REGO` dans Stata qui décompose le R^2 d'un modèle OLS en groupes de prédicteurs puis ensuite en contributions individuelles, comme ceci va être détaillé plus loin. (Hüttner et Sunder (2011 ; 2012)). Ils soulignent l'intérêt en termes de réduction du temps de calcul par l'utilisation de la Owen value sur des groupes par comparaison à la Shapley Value utilisée directement sur l'ensemble des prédicteurs. `REGO` inclut également une option de bootstrap pour calculer des intervalles de confiance. Nous avons également préparé un simulateur avec 3 prédicteurs et une variable à prédire permettant de tester commodément la comparaison des résultats respectifs de la Shapley value et de la Owen value.

Définition : Importance Owen

L'ensemble P des prédicteurs est partitionné en k sous-groupes non vides.

Soit $G(P)$ *une partition de P en k sous- ensembles :* $G(P)=\{G_1,...G_k\}$ *,soit* G_l *un groupe et* X_{jl} *le prédicteur j du groupe* G_l *.*

Une contribution marginale de x_{jl} *ordonnée est dite « respecter la partition* $G(P)$ *» si la coalition S (cf. définition page 49) est telle que pour chaque groupe* $m \neq l$ *soit* $G_m \in S$ *soit* $G_m \cap S = \varnothing$.

Les contributions marginales ordonnées respectant la partition $G(P)$ *forment un sous-ensemble des contributions marginales ordonnées possibles telles qu'elles sont utilisées dans la Shapley Value.*

Soient $\Theta(P,G)$ *L'ensemble des permutations associées à des contributions marginales ordonnées respectant* $G(P)$ *et* $\left|\Theta(P,G)\right|$ *le cardinal de* $\Theta(P,G)$ *. Avec k groupes de cardinaux respectifs* $G_1,...G_k$ *, il y a* $k!$ *permutations possibles des k groupes* $G_1,...G_k$ *et pour chaque groupe de cardinal* $\left|G_l\right|$ *il y a* $\left|G_l\right|!$ *permutations possibles. Il en résulte que :*

$$\left|\Theta(P,G)\right| = k!\left|G_1\right|!...\left|G_i\right|!...\left|G_k\right|!$$

$$\mathrm{Ow}(\mathrm{j})\ (\Delta,\mathrm{G}(\mathrm{P})) = \left(\frac{1}{\left|\Theta(\mathrm{P},\mathrm{G})\right|}\right)\left(\sum_{\pi}\Delta_{\pi}(\mathrm{j})\right) \qquad (3.1.4.1)$$

Une autre manière de présenter Owen Value est de considérer deux étapes :

Etape 1 : effectuer une allocation par Shapley value au niveau des différents Groupes G_l (décomposition inter groupes) en utilisant le R^2 comme fonction de gain

Etape 2 : définir pour chaque sous-ensemble C d'un sous-groupe G_l la valeur de gain $g(C)$ comme l'allocation obtenue à l'étape 1 ci-dessus mais appliquée à la sous-partition des prédicteurs composée des $G_k, k \neq l$ et C. Cette fonction de gain permet en utilisant la Shapley Value d'allouer à chaque prédicteur j au sein de G_l la valeur de Owen telle que définie plus haut. Pour une présentation détaillée voir également Hüttner and Sunder (2011).

Hüttner et Sunder ont appliqué l'Owen value au cas de données salariales en Allemagne (Hüttner et Sunder (2011)) ainsi qu'au cas des données auto.dta (Hüttner et Sunder (2012))

Ils relèvent que les deux décompositions de Shapley et d'Owen peuvent accorder une importance à un prédicteur même si son « true β » est nul et soulignent que la décomposition d'Owen peut être moins intensive en termes de calcul car effectivement elle fait intervenir moins de permutations que les calculs de la Shapley Value en raison précisément des regroupements effectués.

Grâce au simulateur créé et présenté ci-après nous pouvons choisir des structures de données puis calculer et comparer les valeurs de Shapley et d'Owen dans le cas de 3 prédicteurs. Considérons trois prédicteurs a, b, c et une variable à prédire y. a, b, c et y sont ici trois variables aléatoires réelles de variance finie.

Dans ce cas simple avec 3 variables il est possible d'écrire de façon explicite les valeurs de Shapley et Owen en fonction des R^2 théoriques associés à chaque sous-modèle linéaire.

Pour chaque sous ensemble E de l'ensemble formé par les 3 prédicteurs notons $R^2(E)$ le coefficient de détermination de la régression linéaire de la variable à prédire sur les prédicteurs de l'ensemble E. Notons $SV(j)$ et $Ow(j)$ les valeurs de Shapley et d'Owen pour la variable j.

Dans ce cas à trois prédicteurs il vient les équations suivantes :

$$SV(a) = \frac{1}{6}(2R^2(a,b,c) + R^2(a,b) + R^2(a,c) - 2R^2(b,c) + 2R^2(a) - R^2(b) - R^2(c))$$
$$SV(b) = \frac{1}{6}(2R^2(a,b,c) + R^2(b,c) + R^2(b,a) - 2R^2(c,a) + 2R^2(b) - R^2(c) - R^2(a))$$
$$SV(c) = \frac{1}{6}(2R^2(a,b,c) + R^2(c,a) + R^2(c,b) - 2R^2(a,b) + 2R^2(c) - R^2(a) - R^2(b))$$

Si nous considérons les deux coalitions suivantes : $\{a\}$ d'une part et $\{b,c\}$ d'autre part nous pouvons calculer les valeurs de Owen pour $a,b,c.$

$$Ow(a) = \left[\frac{1}{2}(R^2(a,b,c) - R^2(b,c) + R^2(a))\right]$$
$$Ow(b) = \frac{1}{2}\left[\frac{1}{2}(R^2(a,b,c) + R^2(b,c) + R^2(a,b) - R^2(a,c) - R^2(c) + R^2(b) - R^2(a))\right]$$
$$Ow(c) = \frac{1}{2}\left[\frac{1}{2}(R^2(a,b,c) + R^2(b,c) - R^2(a,b) + R^2(a,c) - R^2(b) + R^2(c) - R^2(a))\right]$$

Equations 3.1.4.2. Equations de la décomposition d'Owen dans le cas de 3 prédicteurs.

Ceci permet de créer un simulateur et de comparer les résultats de Shapley Value de Owen Value appliqués à différentes structures choisies de prédicteurs, avec comme partition de Owen : {a} ;{b, c}.

Une structure peut être simulée de la façon suivante :

Soient 3 variables aléatoires centrées réduites *X, Y* et *Z* (par exemple suivant une loi normale).

Une structure peut être générée à partir des 5 paramètres que sont les angles $B,\varphi,C,\psi,\omega.$

- $a = X$
- $b = \cos(B)X + \sin(B)Y$
- $c = \sin(\varphi)\cos(C)X + \sin(\varphi)\cos(C)Y + \cos(\varphi)Z$
- $y = \sin(\psi)\cos(\omega)X + \sin(\psi)\sin(\omega)Y + \cos(\psi)Z$

Les différents choix de paramètres permettent ainsi de générer des structures et d'évaluer les différences entre *lmg-Shapley* et *Owen value.*

Quelques exemples sont présentés ci-après :

Structure 1 :

ρab=	0,799	ρya=	0,889
ρbc=	0,358	ρyb=	0,937
ρca=	0,352	ρyc=	0,601
SV(a)=	0,373	Ow(a)=	0,416
SV(b)=	0,454	Ow(b)=	0,433
SV(c)=	0,172	Ow(c)=	0,151
Σ	1,000	Σ	1,000

Tableau 3.1.4.1. Comparaison entre *lmg-Shapley* et *Owen*. Structure 1.

Voici maintenant un scenario ou a et b sont très corrélés.

Structure 2 :

ρab=	0,999	ρya=	0,889
ρbc=	0,919	ρyb=	0,902
ρca=	0,908	ρyc=	0,999
SV(a)=	0,284	Ow(a)=	0,396
SV(b)=	0,295	Ow(b)=	0,240
SV(c)=	0,421	Ow(c)=	0,365
Σ	**1,000**	**Σ**	**1,000**

Tableau 3.1.4.2 Comparaison entre *lmg-Shapley* et *Owen*. Structure 2.

Comme attendu avec la forte colinéarité entre a et b la différence entre $SV(a)$ et $SV(b)$ tend à se réduire tandis que les Owen Values restent plus étagées, en raison du fait que b « joue » avec c dans la partition retenue ici.

Dans le cas extrême ou b et c sont parfaitement corrélées leurs allocations respectives avec la méthode lmg-Shapley et la méthode d'Owen sont dans les deux cas identiques, mais la répartition relative entre a d'une part et b, c d'autre part est différente comme le montre la structure ci-après :

Structure 3 :

ρab=	0,707	ρya=	0,577
ρbc=	1,000	ρyb=	0,816
ρca=	0,707	ρyc=	0,816
SV(a)=	0,111	Ow(a)=	0,167
SV(b)=	0,444	Ow(b)=	0,417
SV(c)=	0,444	Ow(c)=	0,417
Σ	**1,000**	**Σ**	**1,000**

Tableau 3.1.4.3. Comparaison entre *lmg-Shapley* et *Owen*. Structure 3.

Voici maintenant les résultats de cas étudiés par Hüttner et Sunder(2012) avec les données auto.dta (cf. détails en annexe).

Dans un premier cas étudié mpg (c3) a été modélisée sur 4 prédicteurs :

- weight (c7)
- length (c8)
- headroom (c5)
- price (c2)

Le R^2 est de 0,6668 et Sunder et Hüttner présentent une décomposition de la variance par Shapley Value et Owen Value en utilisant la partition des prédicteurs suivantes : $[\{weight, length\}, headroom, price]$

Les résultats sont résumés ci-après en % :

Variable	Shapley %	Owen % Individuels	Owen % Groupe
weight	41,0462	36,8744	72,2732
length	39,5705	35,3988	
headroom	8,1060	11,8612	11,8612
price	11,2773	15,8656	15,8656
Total	100	100	

Tableau 3.1.4.4. Comparaison *lmg-Shapley, Owen*. Données *auto.dta*.

Comme dans le cas des simulations précédentes les valeurs sont cohérentes entre Shapley Value et Owen Value mais présentent quelques différences dues aux influences des coalitions possibles avec valeur d'Owen.

Dans le cadre d'une discussion directe avec M. Sunder un autre cas a été étudié en utilisant cette fois le jeu auto.dta toujours en regroupant « weight » et « length » mais en ne prenant en compte que la seule variable « price ».

Voici les résultats transmis par M Sunder :

Variable	Shapley %	Owen % Individuels	Owen % Groupe
weight	44,832	42,283	83,145
length	43,410	40,861	
price	11,757	16,855	16,855
Total	100	100	

Tableau 3.1.4.5. Comparaison *lmg-Shapley, Owen*. Données *auto.dta*.

Le R^2 est ici égal à 0,6665 légèrement inférieur à celui observé plus haut avec 4 prédicteurs. Le prédicteur abandonné par rapport au modèle avec 4 prédicteurs, « headroom » est d'ailleurs celui avec la plus faible corrélation avec la variable à prédire (-0,41). Sans surprise « weight » et « length » qui ont une corrélation très forte (0,94) reçoivent dans les deux cas des valeurs très proches à la fois par Shapley et par Owen.

En revanche le prédicteur « price » reçoit une allocation plus forte dans le cas de l'Owen value qui d'une certaine façon « pénalise » le couple très corrélé « weight-length ». A noter que si nous considérons seulement un modèle

avec deux prédicteurs (« price » et « length ») le prédicteur « price » reçoit cette fois 18 % de la part de variance. Ce résultat est à mettre aussi en regard par exemple de la valeur d'allocation *first* pour « price » qui est ici de 14,6 % en pourcentage dans le modèle à 3 variables.

Le fait de regrouper des variables corrélées, ce qui est assez logique, permet ainsi avec la valeur d'Owen de réduire l'inflation de leur part de variance dans le jeu de Shapley. En revanche notons que cet effet n'est pas systématique et que dans certaines configurations particulières de prédicteurs nous pouvons avoir un écart inverse, qui est assez proche de ce qui a été discuté plus haut à propos des « last » et des « first ».

Dans le cas de trois prédicteurs avec la partition $\{a\}$ et $\{b,c\}$ il vient :

$$Ow(a) - SV(a) = \frac{1}{6}\left[R^2(a,b,c) - R^2(a,b) - R^2(a,c) - R^2(b,c) + R^2(a) + R^2(b) + R^2(c) \right] \quad (3.1.4.2)$$

Cette équation peut d'ailleurs être réécrite et utilisant les valeurs F et L, somme des *first* et *last,* ici pour 3 prédicteurs.

$$Ow(a) - SV(a) = \frac{1}{6}\left[F + L - 2R^2 \right] \quad (3.1.4.3)$$

Dans les exemples présentés plus haut le prédicteur isolé (a) gagne en importance relative par rapport aux deux autres « coalisés » (b et c). Mais si les trois prédicteurs sont très corrélés et la variable à prédire très décorrélée de ces prédicteurs la valeur numérique des termes de l'équation (3.1.4.3) peut être négative.

Ceci est le cas dans la configuration suivante :

Avec les notations précédentes :

	degrés
B	21,0
ϕ	85,0
C	10,0
ψ	80,0
ω	77,0

Les résultats sont alors les suivants :

ρab=	0,934	ρya=	0,222

ρbc=	0,978	ρyb=	0,551
ρca=	0,981	ρyc=	0,398
SV(a)=	0,338	Ow(a)=	0,149
SV(b)=	0,379	Ow(b)=	0,473
SV(c)=	0,283	Ow(c)=	0,378
Σ	**1,000**	**Σ**	**1,000**

Tableau 3.1.4.6 ; Comparaisons *lmg-Shapley, Owen*. Données simulées.

Dans cette configurations la valeur de Owen du prédicteur a avec la partition $\{a\},\{b,c\}$ est inférieure à sa valeur de Shapley. Ceci démontre également que dans le cas de 3 prédicteurs $2R^2$ n'est pas un minorant de $F+L$ car il est possible que les deux termes de l'équation 3.1.4.3 soient négatifs.

Les résultats présentés par Sunder dans le cas de trois variables sont totalement identiques à ceux calculés indépendamment dans le cadre de cette recherche en utilisant le package relaimpo pour calculer les décompositions de Shapley et les équations (3.1.4.2) ci-dessus pour les valeurs d'Owen en utilisant les R^2 suivants pour les sous-modèles, calculés avec le modèle linéaire dans R :

$$\begin{aligned} R^2(a,b,c) &= 0,666483 \\ R^2(b,c) &= 0,6613904 \\ R^2(a,b) &= 0,6531447 \\ R^2(a,c) &= 0,6524575 \\ R^2(a) &= 0,2195829 \\ R^2(b) &= 0,6515313 \\ R^2(c) &= 0,6332649 \end{aligned} \qquad (3.1.4.4)$$

Cette égalité de résultats est un élément de confirmation de l'exactitude des équations (3.1.4.2).

Hüttner et Sunder (2011) ont aussi appliqué Shapley et Owen values sur un jeu de données économiques mais qui ne sont pas librement disponibles, et dont les résultats sont présentés en annexe. Notons que ces auteurs ont utilisé l'Owen value pour des regroupements très intuitifs et logiques. Nous n'avons pas trouvé dans la pratique des études de marché d'utilisation de la décomposition d'Owen et cette méthode n'est pas non plus mentionnée dans l'article de synthèse publié par Grömping en mars 2015 (Grömping 2015).

L'utilisation de la décomposition via les valeurs d'Owen ne paraît pas apporter de bénéfice flagrant en termes d'interprétation. En revanche elle est attractive pour réduire les temps de calcul.

Si des variables très corrélées sont mises dans des partitions différentes nous pouvons obtenir des résultats assez différents dans les allocations, tandis que si nous les mettons dans la même partition, ce qui est logique, nous obtenons des valeurs très proches car il s'agit en fait d'une application particulière de la Shapley Value au sein d'un jeu coopératif.

Au total l'utilisation de la décomposition d'Owen suppose une connaissance de groupements *a priori* et dans ce cas il paraît tentant de procéder à une sélection de variables dans les groupes ou de mettre en œuvre un modèle structuré (chemin/SEM).

La décomposition avec l'Owen value, par ses avantages en termes de temps de calculs, et par l'intérêt de procéder à des regroupements, mériterait une place formelle dans les mesures de décomposition de la variance et elle aurait dû être mentionnée par Grömping dans sa synthèse (Grömping (2015)). A ce titre un développement de package pourrait aussi être envisagé.

3.1.5 Décompositions par poids relatifs (Relative Weights Allocations)

Parmi les autres méthodes de décomposition de la variance plusieurs consistent à utiliser des vecteurs et matrices particuliers intervenant dans la décomposition en valeurs singulières. La décomposition de la variance expliquée en utilisant des vecteurs orthogonaux quelconques de l'espace engendré par X peut aussi être généralisée par l'approche de l'allocation pondérée de la variance expliquée (Relative Weights Allocation) présentée ci-après.

Notations :

$\mathbf{X} = \left(x_{ij} \right)_{i=1,n;\, j=1,p}$ est la matrice (n, p) des observations des prédicteurs, supposée de rang p.

$\mathbf{y} = \left(y_i \right)_{i=1,n}$ est le vecteur des observations de la variable à prédire

Dans ce qui suit nous considérons que les valeurs observées sont centrées et réduites c'est-à-dire :

Soit $\mathbf{x_j} = \left(x_{ij} \right)_{i=1,n}$

$\overline{\mathbf{x_j}} = 0$ et $V(\mathbf{x_j}) = 1$ et aussi $\overline{\mathbf{y}} = 0$ et $V(\mathbf{y}) = 1$.

$\mathbf{y^* = Xb = (X'X)^{-1} X'Y}$ est la projection de $\mathbf{y}$ sur l'espace engendré par les vecteurs colonnes de $\mathbf{X}$.

$V(\mathbf{y}^*)$ est la variance expliquée par la régression linéaire.

Définition : Relative Weight Allocation

Une allocation de poids relatifs (Relative Weight Allocation) est effectuée en deux étapes :

Première Etape : Réaliser une allocation de $V(\mathbf{y}^*)$ *sur chaque prédicteur j c'est-à-dire décomposer* $V(\mathbf{y}^*)$ *en p termes positifs* a_j *dont la somme soit égale à* $V(\mathbf{y}^*)$.

$$V(\mathbf{y}^*) = \sum_{j=1}^{j=p} a_j \qquad (3.1.5.1)$$

Notons $\mathbf{A}$ *le vecteur colonne des* a_j.

Deuxième Etape : Chaque prédicteur X_k *se voit allouer une somme pondérée des termes* a_l *avec pour poids* π_{kl} *:*

$$RW(k) = \sum_{l=1}^{l=p} \pi_{kl} a_l \qquad (3.1.5.2)$$

Notons :

$\mathbf{\Pi}$ *la matrice carrée (p,p) des poids et*

$\mathbf{R_w}$ *le vecteur des « Relative Weights Allocations » RW(k):*

$$\mathbf{\Pi} = \begin{pmatrix} \pi_{11} & \cdots & \pi_{1p} \\ \vdots & \ddots & \vdots \\ \pi_{p1} & \cdots & \pi_{pp} \end{pmatrix} \qquad (3.1.5.3)$$

$$\mathbf{R_w} = \begin{pmatrix} RW(1) \\ . \\ RW(p) \end{pmatrix}$$

$$\mathbf{R_w = \Pi A} \qquad (3.1.5.4)$$

Les $RW(j)$ formeront une décomposition de la variance si $\sum_{j=1}^{j=p} RW(j) = V(\mathbf{y}^*)$

Ce qui signifie :

$$\sum_{k=1}^{k=p} RW(k) = \sum_{k=1}^{k=p} \left(\sum_{l=1}^{l=p} \pi_{kl} a_l\right) = \sum_{l=1}^{l=p} a_l \left(\sum_{k=1}^{k=p} \pi_{kl}\right) = V(\mathbf{y}^*) \qquad (3.1.5.5)$$

Compte tenu de l'équation (3.1.5.1) l'équation (3.1.5.5) sera vérifiée si :

$$\forall l, \sum_{k=1}^{k=p} \pi_{kl} = 1 \qquad (3.1.5.6)$$

Si la matrice $\mathbf{\Pi}$ est constituée de vecteurs colonnes unitaires, cette condition (3.1.5.6) est respectée. Cette situation sera rencontrée dans plusieurs des exemples qui suivent et les méthodes de décomposition présentées dans les publications relatives à la décomposition de la variance (Fabbris (1980), Genizi (1993), Johnson (2000)) peuvent être caractérisées par les choix de $\mathbf{\Pi}$ et de $\mathbf{A}$.

En particulier à partir de n'importe quelle base orthonormale $\mathbf{Z}$ dans l'espace engendré par les vecteurs colonnes de $\mathbf{X}$ dans $\mathbf{R}^n$ il est possible de procéder à une « Relative Weights Allocation » de la façon suivante :

- premièrement régression de $\mathbf{y}$ sur des prédicteurs orthogonaux $\mathbf{Z}$,
- deuxièmement régression des $\mathbf{X}$ sur les prédicteurs orthogonaux $\mathbf{Z}$
- troisièmement à procéder à une décomposition de la variance expliquée.

Ceci revient effectivement à effectuer une RWA avec :

$$a_j = r^2(\mathbf{y}, \mathbf{x_j})$$

$$\pi_{ij} = r^2(\mathbf{z_i}, \mathbf{x_j})$$

Pour toute matrice $\mathbf{M}$ de dimensions k, l nous noterons $\mathbf{M}^{\cdot 2}$ son carré de Hadamard, en d'autres termes la matrice formée en élevant chaque terme de $\mathbf{M}$ au carré. Cette matrice $\mathbf{M}^{\cdot 2}$ correspond dans le code R à $\mathbf{M*M}$.

Nous pouvons donc associer à chaque base orthogonale $\mathbf{Z}$ de l'espace engendré par $\mathbf{X}$ dans R^n une allocation de poids relatifs définie par :

$$\mathbf{A = (Z'y)^{\cdot 2}}$$
$$\mathbf{\Pi = (Z'X)^{\cdot 2}}$$

Les termes de la matrice $\mathbf{\Pi}$ respectent la condition (3.1.5.6).

Afin d'éviter l'ambiguïté du terme d'importance relative des prédicteurs, Johnson (2000) a souhaité parler de Poids Relatifs (Relative Weights). Plusieurs méthodes ont été développées selon l'approche des poids relatifs et sont présentées ci-après. Nous monterons plusieurs résultats contestant des conclusions antérieurement publiées et une méthode additionnelle sera également proposée.

Néanmoins il est à ce stade intéressant de synthétiser trois méthodes qui vont être présentées en fonction des matrices et vecteurs de poids utilisés :

	Matrice Π	Vecteur A
Fabbris	$\mathbf{U}^{.2}$	$(\mathbf{V'y})^{.2}$
Genizi-Johnson	$(\mathbf{Z'X})^{.2}$	$(\mathbf{Z'y})^{.2}$
CAR Scores	$\mathbf{I}$	$(\mathbf{Z'y})^{.2}$

La notation $\mathbf{M}^{.2}$ correspond au carré de Hadamard de la matrice $\mathbf{M}$.

$\mathbf{Z}$ et $\mathbf{U}$ sont calculés à partir de la décomposition en valeurs singulières de la matrice des observations des prédicteurs : $\mathbf{X = V\Lambda^{1/2}U'}$ et $\mathbf{Z = VU'}$.

3.1.6 Méthode de Green

Présentée en 1978 cette méthode fait appel à la décomposition en valeurs singulières de $\mathbf{X}$.

Elle est présentée en utilisant les notations suivantes pour la décomposition en valeurs singulières : $\mathbf{X = V\Lambda^{1/2}U'}$, qui sont aussi les notations utilisées par Fabbris (1980).

Définition : Décomposition de Green

Cette méthode consiste en l'utilisation pour le vecteur A des carrés des corrélations entre y et les vecteurs unitaires formés par les colonnes de $\mathbf{Z = VU'}$. *Donc* $a_k = cor^2(\mathbf{Z_k, y})$. *Les* a_k *sont les carrés des termes du vecteur colonne* $\mathbf{Z'y}$.

La matrice des poids $\mathbf{\Pi}$ *est constituée des moyennes pondérées des carrés des coefficients de régression des Z sur les X ainsi que décrit ci-après :*

$$\text{Soit } \mathbf{\Gamma = (X'X)^{-1}X'Z = U\Lambda^{-1}U'U\Lambda^{1/2}V'VU' = U\Lambda^{-1/2}U'}$$

En notant $\Gamma = (\gamma_{lm})$ *chaque poids* π_{jk} *est calculé comme suit :*

$$\pi_{jk} = \frac{\gamma_{jk}^2}{\sum_{i=1}^{i=p} \gamma_{ik}^2} \qquad (3.1.6.1)$$

Cette méthode a été critiquée car elle fait intervenir une régression des vecteurs colonnes de Z sur les X qui sont en général corrélés entre eux, ce qui entraîne que la somme des carrés des coefficients de corrélations de chaque $\mathbf{Z_j}$ sur les $\mathbf{X}$ n'est pas égal à $\left|\mathbf{Z_j}\right|^2$, d'où la nécessité de la pondération au dénominateur de la formule (3.1.6.1).

La décomposition de Green a été calculée avec le jeu de données *swiss 182* afin de comparer le résultat obtenu avec celui présenté par Grömping (2015). Les résultats sont en fait différents.

Variable	Agriculture	Examination	Education	Catholic	Infant Mortality
Green	6,2%	38,7%	45,5%	8,6%	1,0%
Green (Grömping)	2,0%	44,3%	47,5%	5,3%	0,6%

Tableau 3.1.6.1. *Green* : calcul effectué directement avec le script R en annexe.
Green (Grömping) : résultats d'après Grömping (2015). Données « *swiss* 182 ».

Il a été possible de reconstituer les résultats présentés par Grömping (2015), en fait il apparaît que si la normalisation de l'équation (3.1.6.1) ci-dessus est omise, les résultats après normalisation des importances à 100 % sont les suivants :

Variable	Agriculture	Examination	Education	Catholic	Infant Mortality
green Gammasq	2,4%	44,3%	47,5%	5,3%	0,6%

Tableau 3.1.6.2 : % d'allocations de la variance expliquée avec Z'y et allocation des poids en utilisant les carrés de la matrice Γ sans normalisation par colonnes. Données *swiss* 182.

La seule différence avec le résultat de Grömping (2015) est le poids donné à la variable « Agriculture », 2,4 % par rapport à 2,0%. Notons cependant que le total des % présenté par Grömping est 99,7% et que la valeur de 2,0 % est donc vraisemblablement mal reportée.
Aussi pour les données *swiss182* la décomposition de Green avec le calcul présenté plus haut fournit une répartition entre prédicteurs plus proche de *lmg-Shapley*. Le calcul effectué par Grömping (2015) donnait l'impression que les résultats avec la méthode *green* étaient plus nettement différent : *« Green method is somewhat off (and has, in the author's opinion, been justly criticized by Johnson as being flawed* » (Grömping 2015). Les résultats de cette recherche ont été communiqués à U. Grömping qui va procéder à une modification de sa publication.

3.1.7 Méthode de Fabbris

Avec la méthode de Green, la nécessité de normaliser les colonnes de Γ a été critiquée et a conduit Fabbris (1980) à proposer une autre approche. Fabbris utilise également la décomposition en valeur singulière de X :
Soient : $\mathbf{X = V\Lambda^{1/2}U'}$, $\boldsymbol{\beta}^* = \mathbf{V'y}$ et $\mathbf{R_{XV} = X'V = U\Lambda^{1/2}}$

Définition : Décomposition de Fabbris

Fabbris définit w_j *noté ci-après F(j) de la façon suivante :*

$$F(j) = w_j = \sum_{i=1}^{i=p} (\frac{r_{ji}^2}{\lambda_i}) \beta_i^{*2} \qquad (3.1.7.1)$$

r_{ji} étant le coefficient de corrélation entre x_j et v_i , la ième colonne de V.

En remarquant que $\left(\mathbf{R_{XV}}\right)_{lm} = \left(\mathbf{U\Lambda^{1/2}}\right)_{lm} = \left(u_{lm}\lambda_m^{1/2}\right)$, l'équation (2.3.7.1) peut être aussi écrite :

$$F(j) = w_j = \sum_{i=1}^{i=p} u_{ji}^2 \beta_i^{*2} \qquad (3.1.7.2)$$

en notant γ_{ji} le coefficient de régression entre $\mathbf{X_j}$ et $\mathbf{V_i}$, et comme $\mathbf{(X'X)^{-1}X'V = U\Lambda^{-1/2}}$, $\gamma_{ji} = \frac{u_{ji}}{\sqrt{\lambda_i}}$ l'équation (3.1.7.2) peut être aussi écrite :

$$F(j) = w_j = \sum_{i=1}^{i=p} \gamma_{ji}^2 \lambda_i \beta_i^{*2} \qquad (3.1.7.3)$$

Les 3 équations précédentes correspondent aux trois formes écrites dans le texte et dans les notes de l'article de Fabbris (1980).

En reprenant les notations $\mathbf{A}$ et $\mathbf{\Pi}$

- le vecteur $\mathbf{A}$ est constitué des carrés des corrélations entre $\mathbf{y}$ et les vecteurs colonnes de la matrice $\mathbf{V}$,
- et la matrice des poids $\mathbf{\Pi}$ pour allouer les termes carrés des β_i^{*2} selon la formule de Fabbris sont les termes de la matrice $\mathbf{U}$ élevés au carré soit dans la notation de svd v*v.

Le script R est :

```
# betastar correspond à la formule (8) de Fabbris
betastar<-cor(svd$u,y)
betastar2<-betastar*betastar
# Calcul des valeurs de la décomposition de Fabbris avec la formule de "Notes 2" de l'article
de Fabbris
U2<-svd$v*svd$v
Fabbris<-U2%*%betastar2
colnames(Fabbris)<-c("fabbris")
```

Grömping (2015) a indiqué que les méthodes de Fabbris et de Genizi (méthode qui sera présentée au paragraphe suivant) donnaient les mêmes résultats. L'utilisation du script précédent avec le jeu de données *swiss* 182 donne les résultats suivants qui sont comparés à résultat de la méthode *genizi* :

Variable	Agriculture	Examination	Education	Catholic	Infant.Mortality
fabbris	8,5%	41,5%	32,1%	16,6%	1,2%
genizi	6,5%	36,1%	47,7%	8,9%	0,8%

Tableau 3.1.7.1: % d'allocations de la variance explique calculées avec la méthode de Fabbris (1980) et le code ci-dessus, comparée avec les % d'allocations *genizi* calculées avec `relaimpo`. Données *swiss* 182.

Cette comparaison avec la méthode *genizi* montre une différence. En reprenant l'article original de Fabbris (1980), tant les formules théoriques que naturellement les résultats calculés avec R montrent que la méthode de Fabbris n'est pas identique à celle de Genizi-Johnson qui sera présentée plus loin. Sur ce point notre recherche est en contradiction avec les publications de Grömping (2015) et Nimon et al. (2013). Ce résultat a été communiqué à Grömping et sera aussi intégré dans la mise à jour à venir de l'article. Ce point sera en outre complété dans le paragraphe consacré à la méthode de Genizi-Johnson.

Pour le cas simple avec deux prédicteurs, la méthode de Fabbris donnera systématiquement une allocation égale entre les deux prédicteurs ce qui n'est pas le cas avec la méthode *genizi*. En effet, dans le cas de deux prédicteurs, la matrice des poids dans la décomposition de Fabbris est :

$$\mathbf{\Pi} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

Il sera montré au paragraphe suivant que ceci est différent de la décomposition de *genizi/johnson*.

3.1.8 Méthode de Genizi et Johnson

Genizi (1993) et Johnson (2000) ont opté pour une décomposition faisant intervenir la régression de $\mathbf{y}$ et de $\mathbf{X}$ sur les vecteurs $\mathbf{Z}$, avec $\mathbf{Z = VU'}$.

Définition : Décomposition de Genizi et Johnson

En utilisant les notations précédentes, la décomposition de Genizi-Johnson est caractérisée par le vecteur a et la matrice des poids Π suivants :

Le vecteur a et constitué des carrés des corrélations de y avec chaque Z_i *:*

$$a_k = (\mathbf{Z'y})_k^2 = cor^2(\mathbf{Z_k}, \mathbf{y}) \qquad (3.1.8.1)$$

La matrice des poids $\mathbf{\Pi}$ *est :*

$$\mathbf{\Pi} = \left(\pi_{ji}\right) = \left(cor^2(\mathbf{Z_j}, \mathbf{X_i})\right) \qquad (3.1.8.2)$$

Ces vecteurs orthogonaux particuliers issus de la SVD sont proches des $\mathbf{X}$ de départ au sens d'une minimisation de la distance entre les $\mathbf{x_i}$ et les $\mathbf{z_i}$ c'est-à-dire de la somme des carrés des différences entre les $\mathbf{x_i}$ et les $\mathbf{z_i}$ (cf. Johnson (1966)). Cette matrice $\mathbf{Z}$ sera également utilisée par Gibson (1962) puis par Strimmer et Zuber (2011) dans décomposition en « CAR scores » (cf. ci-après).

Notons que de manière générale les résultats d'allocation obtenus dépendent du choix des vecteurs orthogonaux intermédiaires (ci-dessus $\mathbf{Z}$ par exemple). Ainsi une allocation de variance via les vecteurs de $\mathbf{V}$, composantes principales de $\mathbf{XX'}$ ne donnera pas le même résultat que la méthode de Johnson utilisant $\mathbf{Z = VU'}$ (c'est-à-dire en substituant $\mathbf{V}$ à la place de $\mathbf{Z}$ dans les équations (3.1.8.1) et (3.1.8.2).

Nous nous sommes attachés à identifier les correspondances entre la méthode *lmg-Shapley* et celle de Johnson. En effet Johnson et Lebreton (2004) font argument de la proximité des résultats entre la méthode de décomposition proposée par Johnson ceux de *lmg-Shapley* pour en commenter le caractère convaincant :

« *Despite being based on entirely different mathematical models, Johnson's epsilon and Budescu's dominance measures provide nearly identical results when applied to the same data these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors).The convergence between these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors* ."

Nous contestons cette analyse. En vérité le fait que les deux méthodes soient proches ne fournit à notre avis en rien un argument de légitimité. Il est vrai que les résultats de ces deux méthodes dans les exemples publiés sont proches mais ceci est simplement lié au fait que les écarts de variance et la décomposition de Johnson sont liés (comme $1-\cos^2(\alpha)=\sin^2(\alpha)$). En fait dans le cas d'un modèle linéaire avec deux prédicteurs l'allocation de Johnson et la Shapley Value sont strictement identiques.

Ce résultat peut être démontré en utilisant une représentation trigonométrique de la structure des deux prédicteurs et de la variable à prédire. Ceci permet de simuler des configurations variées de la structure des prédicteurs et de la variable à prédire, et de modéliser la décomposition par orthogonalisation en faisant varier les vecteurs orthogonaux utilisés.

Dans le cas de deux prédicteurs nous pouvons formaliser les positionnements des variables et les résultats des décompositions par des formules trigonométriques simples.

Considérons comme au 2.3.1 deux variables aléatoires indépendantes d'espérance nulle de de variance 1 : u et v.

Posons :

$$\mathbf{x_1} = \cos(\varphi)\mathbf{u} - \sin(\varphi)\mathbf{v}$$
$$\mathbf{x_2} = \cos(\varphi)\mathbf{u} + \sin(\varphi)\mathbf{v}$$
$$\mathbf{y} = \cos(\psi)\mathbf{u} + \sin(\psi)\mathbf{v}$$

Par conséquent : $r(\mathbf{x_1}, \mathbf{x_2}) = \cos(2\varphi)$

Soient $\mathbf{z_1}$ et $\mathbf{z_2}$ définis par :

$$\mathbf{z_1} = \cos(\omega)\mathbf{u} + \sin(\omega)\mathbf{v}$$
$$\mathbf{z_2} = -\sin(\omega)\mathbf{u} + \cos(\omega)\mathbf{v}$$

Nous pouvons alors calculer explicitement les valeurs suivantes :

$$SV(1) = lmg(1) = \frac{(1 - \sin(2\psi)\sin(2\varphi))}{2}$$
$$SV(2) = lmg(2) = \frac{(1 + \sin(2\psi)\sin(2\varphi))}{2} \qquad (3.1.8.3)$$

Nous pouvons aussi calculer la décomposition via une base $(\mathbf{z_1}, \mathbf{z_2})$ de vecteurs orthogonaux du plan $(\mathbf{u}, \mathbf{v})$ pour un choix quelconque de ω :

$$VO1 = cos^2(\psi - \omega)cos^2(\omega + \varphi) + sin^2(\psi - \omega)cos^2(\omega - \varphi)$$
$$VO2 = cos^2(\psi - \omega)sin^2(\omega + \varphi) + sin^2(\psi - \omega)sin^2(\omega - \varphi) \qquad (3.1.8.4)$$

Nous avons créé un simulateur permettant une visualisation des résultats et de leur évolution en fonction des paramètres de structure, c'est-à-dire du choix de φ, ψ, ω.

A titre d'exemple dans le graphique ci-dessous nous avons visualisé les évolutions des paramètres β, SV et VO en fonction du choix d'orthogonalisation c'est-à-dire du choix de ω. Les quatre valeurs $\beta_1, \beta_2, lmg(1), lmg(2)$ (rappelons que $SV(j) = lmg(j)$) pour les prédicteurs 1 et 2 sont représentées comme des lignes horizontales et sont constantes car non impactées par le choix de la base d'orthogonalisation c'est-à-dire qu'elles ne dépendent pas de la valeur de ω.

Les valeurs issues de la décomposition orthogonales sont représentées en fonction de l'angle ω entre la première composante principale de $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2})$ et le premier axe orthogonal $\mathbf{z_1}$.

La décomposition de Johnson correspond à $\omega = -\pi/4$, et la décomposition via les composantes principales à $\omega = 0$

.

La configuration des variables $\mathbf{y}$, $\mathbf{x_1}$ et $\mathbf{x_2}$ est caractérisée dans l'exemple suivant par les corrélations suivantes :

$$r(\mathbf{y}, \mathbf{x_1}) = 0,95$$
$$r(\mathbf{y}, \mathbf{x_2}) = 0,48$$
$$r(\mathbf{x_1}, \mathbf{x_2}) = 0,74$$

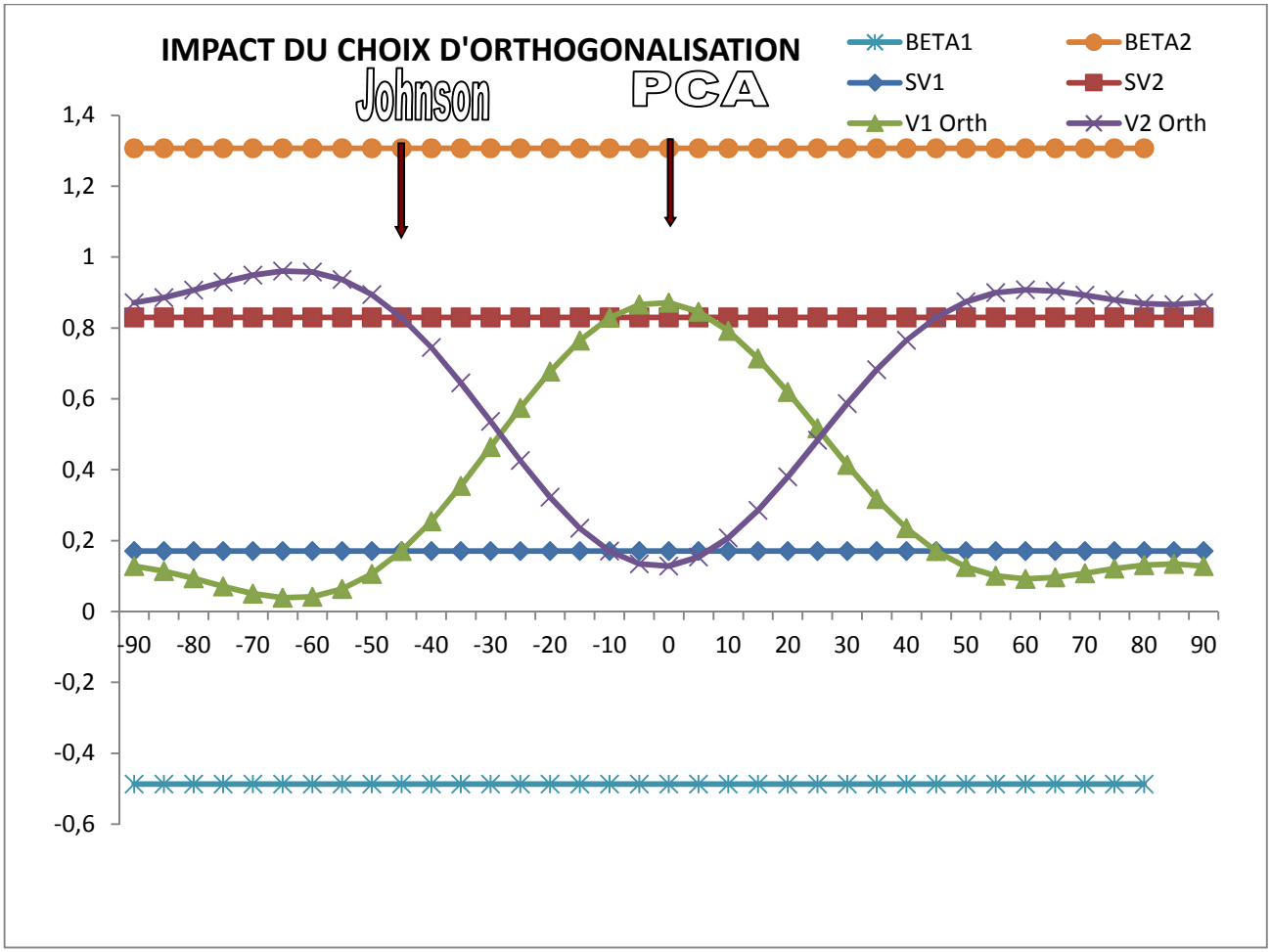


Figure 3.1.8.1. Valeurs de la décomposition orthogonale en fonction du choix de ω. Exemple simulé.

En particulier cet exemple graphique montre l'existence de vecteurs orthogonaux qui permettent via la méthode de décomposition présentée ci-dessus d'allouer aux variables 1 et 2 la même décomposition de variance que la Shapley Value : ces vecteurs orthogonaux sont ceux de la méthode de Genizi-Johnson. C'est en fait le cas général avec deux prédicteurs.

En effet, reportons ω=-π/4 dans les formules (3.1.8.4), il vient :

$$VO(1) = \frac{(1-\sin(2\varphi)\sin(2\psi))}{2}$$

$$VO(2) = \frac{(1+\sin(2\varphi)\sin(2\psi))}{2}$$

Nous reconnaissons dans les expressions de VO(1) et VO(2) les mêmes expressions que SV(1) et SV(2) dans les formules (2.3.8.3).

Donc in fine :

$$VO(1) = SV(1)$$
et
$$VO(2) = SV(2)$$

Cette propriété est aussi obtenue pour le cas où ω=+π/4. Ceci correspond à des $\mathbf{Z}$ orthogonaux qui encadrent le deuxième vecteur propre de $\mathbf{X'X}$.
Ce résultat a été recalculé pour vérification en reprenant les formules de décomposition en valeur singulière de la matrice **X** telles qu'utilisées par Johnson (2000).

Ici :

$$\mathbf{Z'X} = \begin{pmatrix} \cos(-\varphi+\frac{\pi}{4}) & \cos(\varphi+\frac{\pi}{4}) \\ \sin(-\varphi+\frac{\pi}{4}) & \sin(\varphi+\frac{\pi}{4}) \end{pmatrix} \text{ et } \mathbf{Z'y} = \begin{pmatrix} \cos(\psi+\frac{\pi}{4}) \\ \sin(\psi+\frac{\pi}{4}) \end{pmatrix}$$

avec $\mathbf{Z' = P'Q}$ selon les notations de Johnson, ici avec $\mathbf{P' = I}$. $\mathbf{Q}$ est la rotation de π/4 et dans la matrice $\mathbf{Z'}$ les deux colonnes sont effectivement les vecteurs images par $\mathbf{Z'}$ de $\mathbf{x_1}$ (angle –φ avec la première composante principale qui correspond à l'angle 0 degrés) et de $\mathbf{x_2}$ (angle +φ avec la première composante principale qui correspond à l'angle 0 degrés) respectivement.

Et les Relative Weights (cf. Johnson (2000), page 10) sont :

$$\varepsilon_1 = \cos^2(-\varphi+\frac{\pi}{4})\cos^2(\psi+\frac{\pi}{4}) + \cos^2(\varphi+\frac{\pi}{4})\sin^2(\psi+\frac{\pi}{4})$$
$$\varepsilon_2 = \sin^2(-\varphi+\frac{\pi}{4})\cos^2(\psi+\frac{\pi}{4}) + \sin^2(\varphi+\frac{\pi}{4})\sin^2(\psi+\frac{\pi}{4})$$

Un calcul simple montre que les valeurs des deux ε (qui sont les notations de Johnson pour les Relative Weights) sont identiques aux SV calculés plus haut. Ceci vérifie les résultats obtenus en ce qui concerne la stricte égalité entre Relative Weights et Valeurs de Shapley dans le cas de deux prédicteurs.

D'où le résultat suivant :

Résultat : *Dans le cas d'un modèle avec deux prédicteurs les décompositions de la variance via la Shapley Value et via l'orthogonalisation de Johnson (Relative Weights) sont identiques. Cette égalité est aussi vérifiée dans le cas général (c.à.d.* $\mathbf{x_1}$ *et* $\mathbf{x_2}$ *non orthogonaux) pour la décomposition sur les* $\mathbf{z}$ *qui correspondent à* $\omega=+\pi/4$.

Ce résultat est important car il éclaire l'analyse formulée par Johnson et Lebreton sur la proximité entre ces deux décompositions de la variance. L'identité complète dans le cas de deux prédicteurs résulte de propriétés simples des fonctions trigonométriques.

Thomas, Zumbo, Kwan et Schweitzer (2014) mentionnent également ce résultat et soulignent que cette égalité avait été remarquée par Braun and Oswald (2011) en analysant les exemples utilisés par Budescu (1993), et similairement que ce résultat avait été constaté par Shear et al. (2012).

Dans ces conditions nous sommes moins surpris que Johnson de la proximité entre les Shapley values et les valeurs de décomposition Relative Weights. Pour *p* supérieur à deux les résultats entre Shapley value et RWA restent proches ainsi qu'observé par Johnson.

La démonstration de l'équivalence stricte entre Shapley Value et Johnson dans le cas de deux prédicteurs paraît conforter l'analyse selon laquelle y compris dans le cas de plus de deux prédicteurs Shapley value et Johnson auront tendance à donner des répartitions proches, mais ceci ne nous paraît en aucun cas, contrairement aux commentaires de Johnson et Lebreton rappelés plus haut constituer une justification en soi de la validité de ces décompositions particulières de la variance.

Ainsi il paraît injustifié de conclure que "*The convergence between these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors*".

Nous savons déjà que dans le cas de deux prédicteurs les valeurs de Shapley et les Relative Weights sont strictement identiques pour chacun des prédicteurs.

Dans le cas de plus de deux prédicteurs nous pouvons identifier des valeurs atteintes pour les Relative Weights dans différents cas particuliers qui sont les « last » et les « first » des prédicteurs. Pour cela nous allons considérer des bases orthogonales particulières et pratiquer la décomposition associée à ces bases.

Nous étudierons deux cas :

Cas numéro 1 :

Considérons un prédicteur j et une base orthonormée telle que :

$$\mathbf{z_j} = \frac{\mathbf{y^*}}{\left\|\mathbf{y^*}\right\|}$$

Alors $\forall i \neq j, \mathrm{cov}(\mathbf{y}, \mathbf{z_i}) = 0$ *et* $\mathrm{cov}(\mathbf{y}, \mathbf{z_j}) = 1$

$$Donc\ RW(\mathrm{j}) = \sum_{i=1}^{i=p} \lambda_{ji}^2 \beta_j^2$$

avec $\lambda_{ij} = \mathrm{cov}(\mathbf{z_j}, \mathbf{X_i})$ *and* $\beta_i = \mathrm{cov}(\mathbf{y}, \mathbf{z_i})$ *seul* $\beta j \neq 0; \beta j = 1$ *et donc*

$$RW(\mathrm{j}) = \mathrm{cov}^2(\mathbf{y}, \mathbf{x_j}) = \mathrm{first(j)}$$

Cas Numéro 2 :

Considérons cette fois une base orthogonale telle que :

$$\mathbf{z_j} = \frac{\mathbf{u_j}}{\left\|\mathbf{u_j}\right\|}$$

$\mathbf{u_j}$ étant le résidu de la régression de $\mathbf{X_j}$ sur les autres variables de $\mathbf{X}$

Alors dans ce cas :

$$\lambda_{ji} = \mathrm{cov}(\mathbf{z_j}, \mathbf{x_i}) = 0 \text{ si } i \neq j$$

et

$$\lambda^2_{jj} = \mathrm{cov}^2(\mathbf{z_j}, \mathbf{X_j}) = 1 - \mathrm{R}_j^2$$

Aussi :

$\beta_j^2 = \mathrm{cov}^2(\mathbf{y}, \mathbf{z_j})$ qui dans ce cas particulier n'est autre que le carré du coefficient de $\mathbf{x_j}$ dans la régression de $\mathbf{y}$ sur les $\mathbf{X}$.

Donc :

$$RW(\mathrm{j}) = last(j) = b_j^2 Tol(j)$$

En effet le terme de droite est le carré du coefficient de $\mathbf{x_j}$ de la régression de $\mathbf{y}$ sur les $\mathbf{X}$ multiplié par la tolérance de j.

Ces exemples confirment que des Relative Weights générés par les bases orthogonales de décomposition est assez logiquement « proche » des SV même dans le cas de plus de 2 variables.

Revenons maintenant sur les relations entre $last(i), SV(i)$ et $first(i)$ .En utilisant les notations trigonométriques présentées antérieurement nous pouvons calculer dans les cas de deux prédicteurs l'ensemble des valeurs utiles en fonction des angles choisis (cf. Annexe 3).

Nous pouvons désormais reprendre les analyses en ce qui concerne les commentaires de Tabachnick et Lidell sur les relations entre les *last et* les *first.*

La condition $\cos(2\psi)\cos(2\varphi) \leq 0$ (*Condition (C))* entraîne que les *last* deviennent inférieurs au *first* pour chaque prédicteur.

Rappelons que $\rho_{12} = \cos(2\varphi)$) et notons également que tant que cette condition C n'est pas respectée on a bien l'inégalité inverse :

$$last(j) \leq SV(j) \leq first(\text{j})$$

Cette condition caractérise et quantifie sur un exemple simple avec deux prédicteurs les cas que Tabachnick qualifie de configurations très particulières d'inversion d'ordre entre la somme des last, le R² c'est-à-dire comme vu précédemment la somme des SV (ou RW) et la somme des firsts.

Nous pouvons illustrer les variations des last et first pour les deux prédicteurs en faisant varier le coefficient de corrélation entre les deux prédicteurs (variations de φ) tandis que ψ c'est-à-dire in fine le coefficient de corrélation de y avec la première composante principale est maintenu constant (ici ψ= 19 degrés, en abscisse est reporté le coefficient de corrélation entre $\mathbf{x_1}$ et $\mathbf{x_2}$). Dans cette configuration cos (2ψ) est positif et le signe de la condition « C « est donc celui de cos (2φ) c'est-à-dire le signe de ρ_{12} .

L'analyse graphique ci-après confirme que dès que $\mathbf{x_1}$ et $\mathbf{x_2}$ ont une corrélation négative, $last(i)$ devient supérieur au $first(i)$ pour chaque prédicteur.

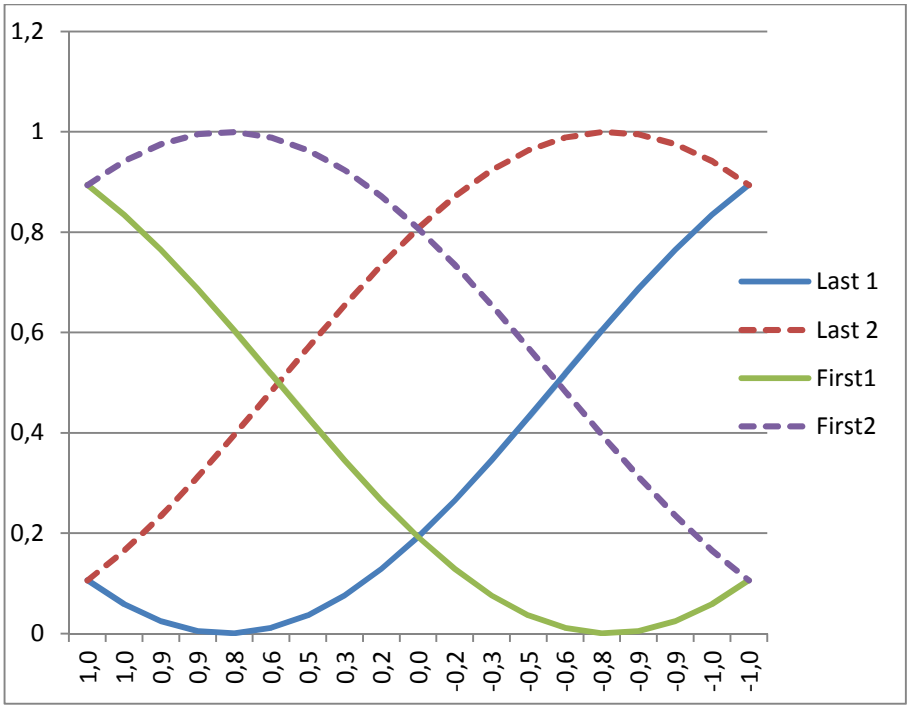


Figure 3.1.8.2. Variation des *last* et *first* de deux prédicteurs en fonction de leur corrélation.

Le graphique suivant présente la comparaison entre somme des *last* et somme des *first* en additionnant les valeurs pour les deux prédicteurs :

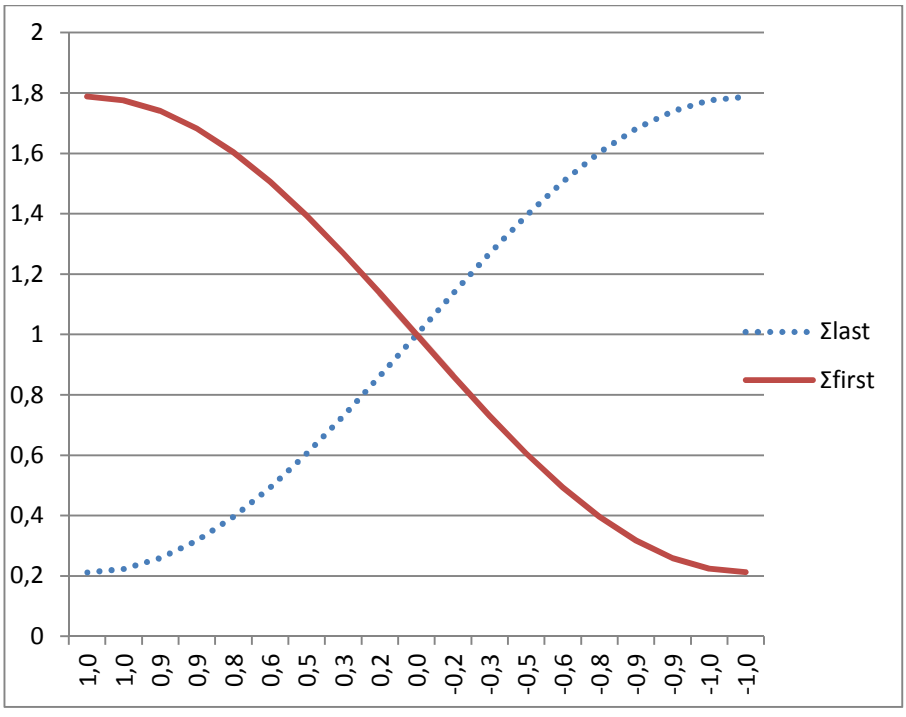


Figure 3.1.8.3. Somme des *last* et des *first* de deux prédicteurs en fonction de leur corrélation.

Comme les valeurs *last* sont en fait les carrés des coefficients de corrélation semi partiels, nous pouvons exprimer la condition précédente sous une autre forme en notant sr_j le coefficient de corrélation semi-partielle :

$$R_y^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2\rho_{12} r_{y1} r_{y2}}{1-\rho_{12}^2}$$

$$R_y^2 = \frac{sr_1^2 + sr_2^2 + 2\rho_{12} sr_1 sr_2}{1-\rho_{12}^2}$$

Nous pouvons écrire dans le cas de deux prédicteurs la condition pour que la somme des *last* soit supérieure au R^2 ² ou de façon équivalente que la somme des *first* soit inférieure au R^2 (puisque dans le cas de deux prédicteurs la somme des *last* plus la somme des *first* est constante) de deux manières équivalentes explicitées ci-après :

Résultat :

Soient deux prédicteurs $\mathbf{x_1}$ *et* $\mathbf{x}_2$ e*t une variable à prédire* $\mathbf{y}$ *:*

Pour que $\sum_j last(j) \le R_y^2 \le \sum_j first(j)$

Il faut et il suffit que :

$$R_y^2 \rho_{12}^2 + 2\rho_{12}^2 sr_1 sr_2 \ge 0 \qquad (3.1.8.5)$$

Ou de façon équivalente que :

$$R_y^2 \rho_{12}^2 - 2\rho_{12} r_{y1} r_{y2} \le 0 \qquad (3.1.8.6)$$

Il facile de vérifier que dans le cas où $\mathbf{y} = \mathbf{y}^*$ ou de manière équivalente en utilisant la projection de $\mathbf{y}$ sur le plan des prédicteurs dans la régression, nous pouvons utiliser la représentation trigonométrique ci-dessus.

La première condition (3.1.8.5) s'écrit alors :

$$\cos(2\varphi)\cos(2\psi) \ge 0$$

Et la deuxième condition (3.1.8.6) :

$$\cos(2\varphi)\cos(2\psi) \ge 0$$

Ces calculs sont cohérents avec les résultats antérieurs c'est à dire que la condition C équivaut dans le cas de deux prédicteurs à l'inversion de l'inégalité (3.1.8.5) ou (3.1.8.6).

Aussi les formules ci-dessus montrent que dans le cas de deux prédicteurs la somme des last plus la somme des firsts est exactement égale à $2R^2$.

Soit plus formellement :

$$last(1) + last(2) + first(1) + first(2) = 2R^2$$

Ce résultat sera repris plus loin sous la forme d'une inégalité caractéristique dans le cas de plus de deux prédicteurs.

Par conséquent il est maintenant immédiatement possible de construire des configurations de variables ne respectant pas l'inégalité « usuelle » au sens ou le disent Grömping, Tabachnick et Fidell, entre :

- la somme des carrés des corrélations semi-partielles,
- le R^2
- et la somme des carrés des corrélations bivariées.

Il suffit de partir de deux variables dé corrélées telles **u** et **v** comme précédemment et de créer les variables $\mathbf{x_1}$, $\mathbf{x_2}$ et **y** en respectant la condition C. Ceci montre qu'il existe toujours des couples de prédicteurs satisfaisant les conditions ci-dessus.

Dans le cas de plus de deux prédicteurs il est par conséquent possible de rencontrer des situations ou la somme des last est supérieure au R^2 , il suffit de considérer le cas de deux prédicteurs dans une configuration comme précédemment c'est-à-dire respectant la condition C , et choisir comme autres prédicteurs des variables deux a deux dé corrélées et également dé corrélées avec les deux prédicteurs particuliers considérés pour in fine générer des variables respectant la conditions C sur deux dimensions parmi les p dimensions. Donc cette situation d'inversion de l'inégalité entre la somme des *last* et la somme des *first* peut apparaître quel que soit le nombre de prédicteurs. D'où le résultat suivant :

Résultat :

Soit une variable à prédire **y** *et p prédicteurs linéairement indépendants, (p supérieur ou égal à 2) il existe des combinaisons linéaires de ces prédicteurs telles que les nouveaux prédicteurs ainsi engendrés soient indépendants et :*

$$\sum last(j) \geq R^2 \geq \sum first(j)$$

Ceci précise les analyses de Grömping, Tabachnick et Fidell évoquées plus haut.

Sur le même sujet et dans le cas de deux prédicteurs, Cohen et al. (1975) se sont eux aussi intéressés (chapitre 3.3.2) au sujet suivant « *Semipartial Correlation Coefficients and Increments to R²* ». Ils relèvent bien que la différence qu'ils notent « c » n'est pas toujours positive.

$$c = r_{12}^2 - sr_1^2 - sr_2^2$$

Ils proposent d'ailleurs un exemple de prédicteurs avec précisément comme indiqué plus haut deux variables apparemment contradictoires pour un même individu (c'est-à-dire qu'en général un individu a des valeurs fortes sur une ou l'autre de ces variables mais rarement sur les deux) mais avec la propriété que des deux prédicteurs ont chacun un impact positif sur la variable à prédire.

Voici la description de cet exemple :

- X1 : Social Assertiveness
- X2 Record Keeping
- Y : Sales Success
- $r_{1y} = 0,403$, $r_{2y} = 0,127$, $r_{12} = -0,305$

Nous trouvons effectivement bien des valeurs telles que la somme des *last* est supérieure à la somme des *firsts* :

Ry1	0,2250
RY2	0,8387
R12	0,7193
β1	0,4870
β2	0,2755
ry1²	0,1624
ry2²	0,0161
Σryi²	0,1785
sr1²	0,2151
sr2²	0,0689
Σsri²	0,2840
Σryi²-Σsri²	-0,1055
R2=	0,2313

Tableau 3.1.8.1 Exemple de configuration ou la somme des *last* est supérieure au R^2

Et la structure peut être représentée graphiquement comme ci-après :

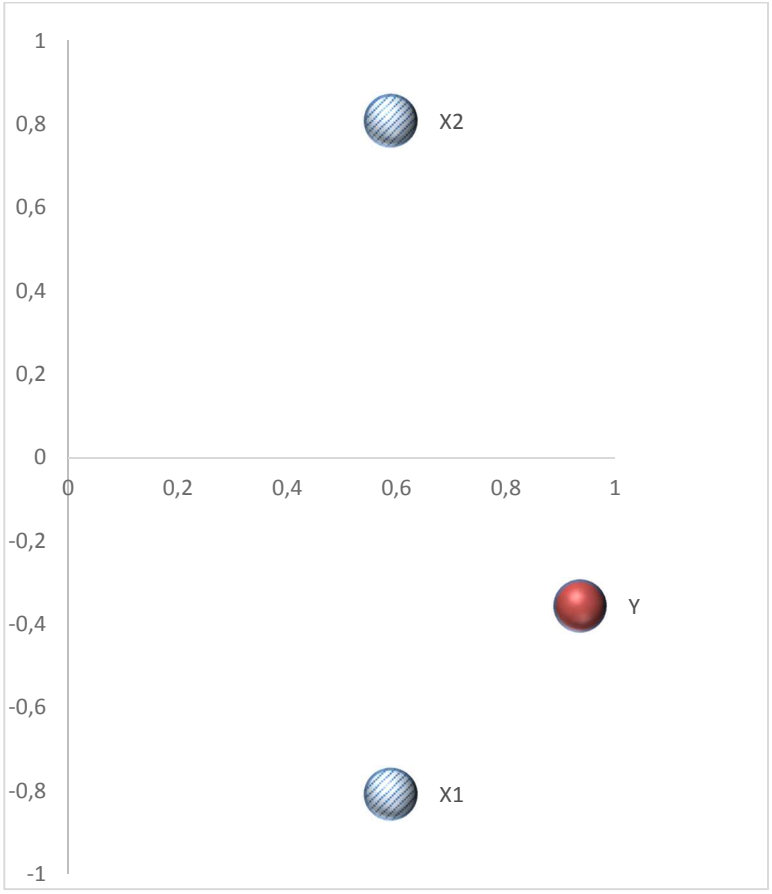


Figure 3.1.8.4

Nous sommes bien dans le cas de la condition C avec cos (2φ) négatif mais cos (2ψ) est ici positif (ψ=-20,8 degrés). Dans cet exemple proposé par Cohen and al $\varphi = 53,88$ degrés.

Dans une édition de 1975 du même ouvrage un autre exemple de structure avait été utilisé (ce texte d'une édition de 1975 est copié dans des notes de cours publiées sur internet)

Structure :

Ry1	0,29
RY2	0,24
R12	-0,3

En conséquence en utilisant le tableur préparé nous trouvons cette fois les valeurs clé suivantes, cohérentes avec le texte de 1975 :

Ry1	0,2250
RY2	0,8387
R12	0,7193
β1	0,3978
β2	0,3593
$ry1^2$	0,0841
$ry2^2$	0,0576
Σryi^2	0,1417
$sr1^2$	0,1440
$sr2^2$	0,1175
Σsri^2	0,2615
$\Sigma ryi^2-\Sigma sri^2$	-0,1198
R2=	0,2016

Tableau 3.1.8.2

Voici une représentation graphique de la structure :

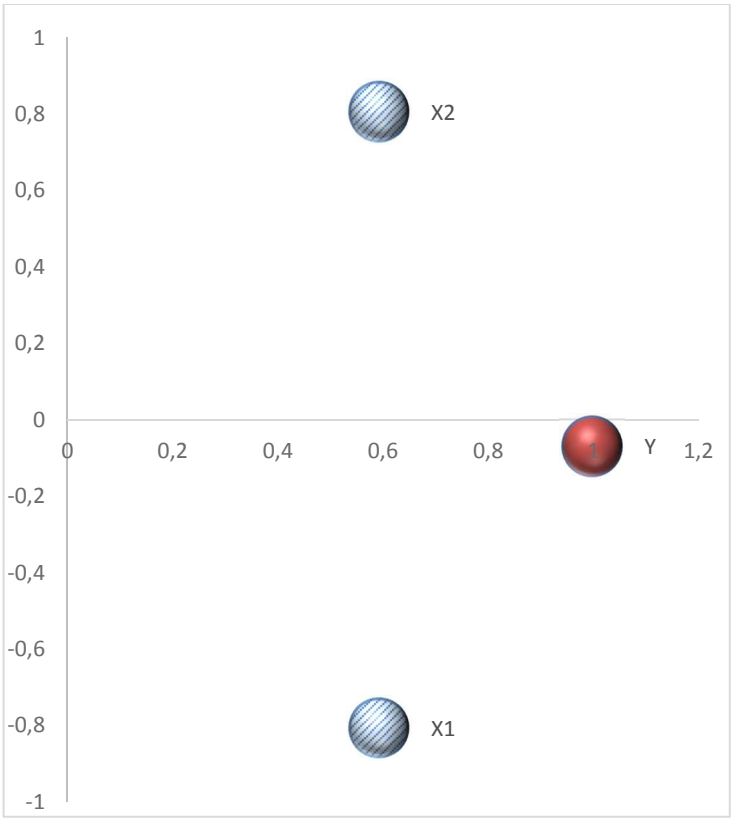


Figure 3.1.8.5

Dans cette configuration φ =53,73 degrés et ψ=-3,96 degrés et la condition « C » est aussi respectée. Mais alternativement le cas où les deux prédicteurs sont fortement corrélés et ou la variable à prédire est décorrélée des deux prédicteurs n'est pas identifié par Cohen et al. (1975). Il s'agit d'un cas a priori assez théorique. Cette configuration est néanmoins parfaitement envisageable si des prédicteurs commodes à mesurer mais corrélés sont utiliser pour prédire une variable en fait assez mal corrélée avec ces mêmes prédicteurs. En voici un exemple simulé ou nous avons également une situation ou la somme des last dépasse le R^2 :

Ry1	0,2250
RY2	0,8387
R12	0,7193
β1	0,7830
β2	1,4019
$ry1^2$	0,0506
$ry2^2$	0,7038
Σryi^2	0,7545
$sr1^2$	0,2961
$sr2^2$	0,9494
Σsri^2	1,2455
$\Sigma ryi^2-\Sigma sri^2$	-0,4910
R2=	1,0000

Tableau 3.1.8.3.

Et en représentation graphique :

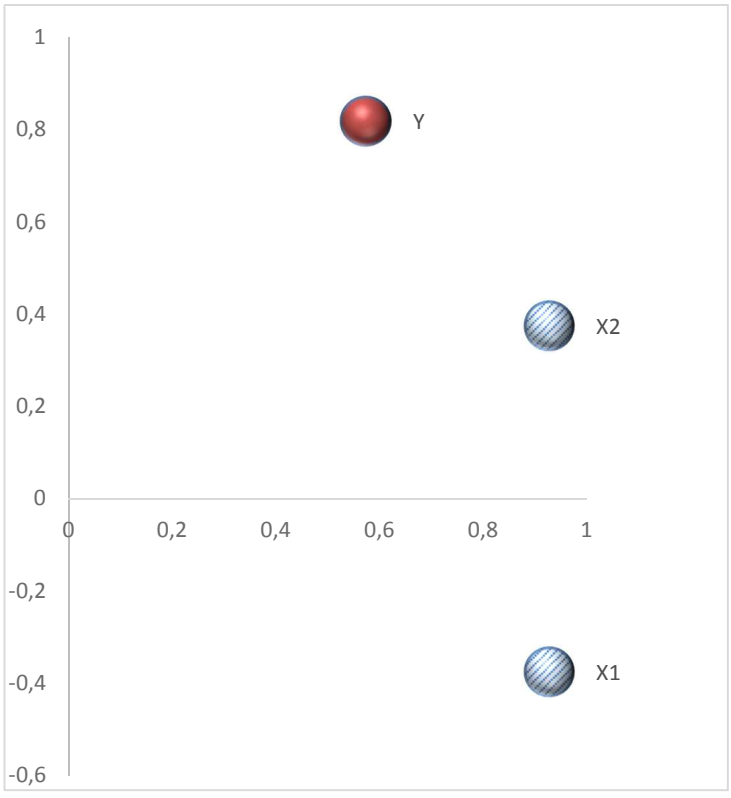


Figure 3.1.8.3

Dans cette configuration φ =22,01 degrés et ψ=54,98 degrés. Ici aussi la condition C est respectée mais cette fois c'est cos (2ψ) qui est le terme négatif.

Au total nous pouvons constater que la décomposition de la variance consiste, qu'il s'agisse de la Shapley value ou de l'orthogonalisation, à déterminer des valeurs d'importance relative par prédicteur qui sont « entre » entre les *last* et les *firsts*.

D'ailleurs dans les cas particuliers à deux variables que nous avons créés comme ci-dessus nous pouvons noter que même si les *last* peuvent devenir supérieurs aux *first* le R^2 est toujours situé dans l'intervalle entre les deux.

En ce qui concerne la décomposition de Johnson-Genizi, Genizi (1993) a voulu comparer les importances obtenues selon plusieurs méthodes : « Relative Weights » c'est-à-dire Genizi avec $\mathbf{Z}$ comme vecteurs orthogonaux, *first, betasquare*, et *lmg*. Il fonde sa comparaison sur l'analyse de la corrélation obtenue entre les séries d'importances entre Relative Weights (Genizi) et les autres méthodes.

Dans le cas de deux prédicteurs il souligne : « *For p=2 all corrélation coefficients were 1, indicating a complete agreement between all measures concerning the order of importance of the two regressors »* (page 418 du document en référence Genizi (1993). Ce résultat est sans réelle portée car la corrélation entre deux variables analysée sur deux observations est toujours 1 : par deux points passe toujours une droite…

3.1.9 Méthode des CAR scores

Cette méthode a été proposée par Gibson (1962) et reprise par Zuber et Strimmer (2011). Elle consiste à allouer à chaque prédicteur j une importance qui est le carré de la corrélation entre la variable à prédire et le $\mathbf{z}_j$ correspondant aux vecteurs de $\mathbf{Z = PQ'}$ de la décomposition de Johnson.

C'est-à-dire en reprenant les notations précédentes :

Définition : carré des CAR scores

$$CAR^2(j) = \beta_j^{*2} = \mathrm{cov}^2(\mathbf{y}, \mathbf{z_j}) \tag{3.1.9.1}$$

Strimmer et Zuber ont appliqué cette décomposition dans le cadre d'analyse multidimensionnelle dans les applications de génétique et biologie moléculaire. Elle est nommée CAR pour **C**orrélation-**A**diusted (marginal) co**R**elation.

L'utilisation directe des coefficients de corrélation avec les vecteurs Z est simple mais elle présente comme inconvénient que les valeurs CAR peuvent rester identiques pour deux prédicteurs même si la corrélation entre ces deux prédicteurs varie. C'est d'ailleurs un point où les résultats présentés par Strimmer et Zuber sont erronés et ces conclusions non fondées sont analysées ci-après.

Il est utile à ce stade de formaliser complètement les matrices utilisées ce qui permettra de montrer pourquoi il faut rejeter l'analyse de Strimmer et Zuber. Soient dans le plan euclidien (O,**u,v**):

$$\mathbf{X} = \begin{pmatrix} \cos(\alpha_1) & \cos(\alpha_2) \\ \sin(\alpha_1) & \sin(\alpha_2) \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} \cos(\frac{\alpha_1+\alpha_2}{2}) & -\sin(\frac{\alpha_1+\alpha_2}{2}) \\ \sin(\frac{\alpha_1+\alpha_2}{2}) & \cos(\frac{\alpha_1+\alpha_2}{2}) \end{pmatrix}$$

$$\mathbf{Q} = \begin{pmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{pmatrix}$$

$$\mathbf{\Delta} = \begin{pmatrix} \sqrt{(1+\rho)} & 0 \\ 0 & \sqrt{(1-\rho)} \end{pmatrix}$$

Avec les notations précédentes où **u** et **v** sont deux variables aléatoires réelles indépendantes et standardisées, chaque point peut ainsi représenter une variable aléatoire combinaison linéaire de **u** et **v**. En choisissant comme axe d'origine la première composante principale $\frac{\alpha_1+\alpha_2}{2} = 0$ et P=I. Aussi $\rho_{12} = \cos(\alpha_2 - \alpha_1) = \cos(2\alpha_2)$

Voici ci-après quelques points présentés par Strimmer et Zuber (2011) et les commentaires qu'ils appellent et qui ont été communiqués à ces deux auteurs.

Commentaire 1

L'article de Strimmer et Zuber (2011) indique (page 12 (4.9)) que lorsque la corrélation entre deux prédicteurs tend vers 1 leur CAR scores tendent à s'égaliser :

"This can directly be seen from the definiion $\omega = P^{1/2} b_{std}$ *of the CAR score. For two predictors* X_1 *and* X_2 *correlation* $Cor(\mathbf{X_1}, \mathbf{X_2}) = \rho$ *a simple algebraic calculation shows that the difference between the two CAR scores equals*

$$\omega_1^2 - \omega_1^2 = ((b_{std})_1^2 - (b_{std})_2^2)\sqrt{1-\rho^2}$$

Therefore, the two squared CAR scores become identical with growing absolute value of the correlation between the variables. This grouping property is intrinsic to the CAR score itself and not a property of an estimator"

Cette assertion est erronée. En réalité $\omega_1^2 - \omega_2^2$ ne dépend que de la correlation entre le prédicteur et la première composante principale de $\mathbf{X = (x_1, x_2)}$. Ceci peut être vérifié en utilisant l'approche trigonométrique pour représenter les prédicteurs $\mathbf{x_1}$ et $\mathbf{x_2}$ avec une corrélation donnée $r(\mathbf{x_1}, \mathbf{x_2}) = r_{12}$ et **y*** étant le résultat de la régression de **y** sur **X**, toutes les variables étant ici centrées et réduites.

$\mathbf{x_1}$, $\mathbf{x_2}$ et y peuvent être exprimés en fonction des deux composantes principales $\mathbf{e_1}$ et $\mathbf{e_2}$ et les *CAR scores* s'en déduisent :

$$\mathbf{X_1} = \cos(\varphi)\mathbf{E_1} - \sin(\varphi)\mathbf{E_2}$$
$$\mathbf{X_2} = \cos(\varphi)\mathbf{E_1} + \sin(\varphi)\mathbf{E_2}$$
$$\mathbf{Y^*} = \cos(\psi)\mathbf{E_1} + \sin(\psi)\mathbf{E_2}$$
$$b_1 = \frac{\sin(\varphi - \psi)}{\sin(2\varphi)}$$
$$b_2 = \frac{\sin(\varphi + \psi)}{\sin(2\varphi)}$$
$$r_{12} = \cos(2\varphi)$$
$$\omega_1 = \cos(\psi + \frac{\pi}{4})$$
$$\omega_2 = \cos(\psi - \frac{\pi}{4})$$

Ceci confirme que :

$$\omega^2_1 - \omega^2_2 = (b^2_1 - b^2_2)\sqrt{1 - r_{12}^2} = -2\sin(2\psi) \qquad (3.1.9.1)$$

Ainsi la différence entre les deux CAR scores ne dépend donc que de ψ, et par conséquent ne dépend pas du degré de corrélation entre $\mathbf{x_1}$ et $\mathbf{x_2}$ mais seulement de la corrélation entre **y** (ou **y***) et $\mathbf{x_1 + x_2}$ (ou pour prendre un vecteur standardisé $\frac{\mathbf{x_1 + x_2}}{\sqrt{2(1 + r_{12})}}$).

A la condition que deux couples de variables $\mathbf{x_1}$ et $\mathbf{x_2}$ et $\mathbf{x'_1}$ et $\mathbf{x'_2}$ partagent la même "bissectrice" les *CAR scores* seront donc identiques.

Soient donc $\mathbf{y}$, $\mathbf{x_1}$ et $\mathbf{x_2}$ donnés avec $\mathbf{x_1}$ et $\mathbf{x_2}$ ayant une corrélation non nulle strictement positive.

Nous pouvons construire $\mathbf{x_1'}$ et $\mathbf{x_2'}$ de la façon suivante :

$$\mathbf{x_1'} = \cos(\varphi')\left[\frac{\mathbf{x_1 + x_2}}{\sqrt{2\sqrt{(1+\rho)}}}\right] - \sin(\varphi')\left[\frac{\mathbf{-x_1 + x_2}}{\sqrt{2\sqrt{(1-\rho)}}}\right]$$

$$\mathbf{x_2'} = \cos(\varphi')\left[\frac{\mathbf{x_1 + x_2}}{\sqrt{2\sqrt{(1+\rho)}}}\right] + \sin(\varphi')\left[\frac{\mathbf{-x_1 + x_2}}{\sqrt{2\sqrt{(1-\rho)}}}\right]$$

La corrélation entre $\mathbf{x_1'}$ et $\mathbf{x_2'}$ est $\rho' = \cos(2\varphi')$ et les *CAR scores* associés sont égaux à ceux générés avec $\mathbf{x_1}$ et $\mathbf{x_2}$.

Donc en faisant tendre φ' vers zéro on peut avoir une corrélation croissante entre $\mathbf{x_1'}$ et $\mathbf{x_2'}$ sans que *les CAR scores* s'égalisent, ce qui contredit l'assertion de Strimmer et Zuber.

Ci-après le même calcul en utilisant des notations algébriques et non plus trigonométriques, avec C_1 et C_2 les deux composantes principales et ω_1 et ω_2 les deux *CAR scores*.

$$\mathbf{C_1} = \frac{\mathbf{x_1 + x_2}}{\sqrt{2}\sqrt{(1-\rho)}}; \mathbf{C_2} = \frac{\mathbf{x_2 - x_1}}{\sqrt{2}\sqrt{(1-\rho)}}$$

$$\mathbf{\Omega_1} = \frac{\sqrt{2}}{2}(\mathbf{C_1 - C_2}); \mathbf{\Omega_2} = \frac{\sqrt{2}}{2}(\mathbf{C_1 + C_2})$$

$$\omega_1^2 - \omega_2^2 = r_{y^*\Omega 1}^2 - r_{y^*\Omega 2}^2 = -2(r_{yC1} r_{yC2})$$

$$(r_{yC1} r_{yC2}) = 2\sin(\psi)\cos(\psi)$$

Nous avons contacté Strimmer et Zuber sur ces questions et M. Strimmer nous a adressé un document explicatif et un script R pour présenter ses résultats avec un exemple numérique destiné à justifier l'assertion mentionnée ci-dessus. Ce code est joint en annexe. Ce code considère une variable à prédire $\mathbf{y}$ définie à partir des deux coefficients notés sbeta1 et sbeta2 appliqués à deux prédicteurs standardisés $\mathbf{x_1}$ et $\mathbf{x_2}$ et effectue une série de calculs en faisant varier le coefficient de corrélation entre les deux prédicteurs. Nous avons fait remarquer à Zuber et Strimmer que leur script revient en fait à faire varier la variable à prédire $\mathbf{y}$ en même temps que les deux prédicteurs alors qu'en fait pour analyser l'impact de la corrélation croissante entre $\mathbf{x_1}$ et $\mathbf{x_2}$ il faut considérer un fixe (appelé ci-après $\mathbf{y_0}$). Et en fait nous avons montré que l'écart entre les *CAR scores* de $\mathbf{x_1}$ et $\mathbf{x_2}$ restait même en fait constant dans certaines configurations bien choisies comme dans la formalisation trigonométrique utilisée au paragraphe précédent et donc ne tendait pas vers zéro même si le coefficient de corrélation entre les deux prédicteurs tendait vers 1.

Ceci est illustré ci-après avec une comparaison entre les *CAR scores* appliqués dans deux cas qui doivent être bien distingués :

Soit les *CAR scores* sont appliqués à une variable à prédire $\mathbf{y_0}$ fixe qui correspond aux variables de départ de l'exemple de Zuber, c'est-à-dire tel que : $b_1 = 0,6$, $b_2 = 0,3$ et $\rho = 0,5$. Nous ferons ensuite varier $\mathbf{x_1}$ et $\mathbf{x_2}$, en changeant leur corrélation et donc bien sur les b_1 et b_2 coefficients de la régression de $\mathbf{y_0}$ fixe sur les nouveaux $\mathbf{x_1}$ et $\mathbf{x_2}$ seront différents des valeurs initiales.

Soit les *CAR scores* sont appliqués à une variable à prédire $\mathbf{y}$ définie comme la variable qui correspond aux sbeta1 et sbeta2 initiaux (c'est-à-dire 0,6 et 0,3 respectivement), mais appliqués aux nouveaux $\mathbf{x_1}$ et $\mathbf{x_2}$ correspondant en réalité à une nouvelle corrélation entre les prédicteurs.

Voici un exemple numérique comparant le cas de référence de Zuber et un autre couple de $\mathbf{x_1}$ et $\mathbf{x_2}$.

Cas de référence : $b_1 = 0,6 \quad b_2 = 0,3 \quad r_{12} = 0,5$

Nous pouvons écrire complètement les vecteurs dans les bases $\mathbf{x_1}$ et $\mathbf{x_2}$ et aussi dans $\mathbf{e_1}$ et $\mathbf{e_2}$:

$$\mathbf{y_0} = 0,6\mathbf{x_1} + 0,3\mathbf{x_2} = 0,78\mathbf{e_1} - 0,15\mathbf{e_2}$$
$$\mathbf{x_1} = 0,87\mathbf{e_1} - 0,5\mathbf{e_2}$$
$$\mathbf{x_2} = 0,87\mathbf{e_1} + 0,5\mathbf{e_2}$$

Autre cas : $b_1 = 0,6 \quad b_2 = 0,3 \quad r_{12} = 0,8$

$$\mathbf{y_0} = 0,648\mathbf{x_1} + 0,174\mathbf{x_2} = 0,779\mathbf{e_1} - 0,150\mathbf{e_2}$$
$$\mathbf{y} = 0,6\mathbf{x_1} + 0,3\mathbf{x_2} = 0,854\mathbf{e_1} - 0,094\mathbf{e_2}$$
$$\mathbf{x_1} = 0,95\mathbf{e_1} - 0,32\mathbf{e_2}$$
$$\mathbf{x_2} = 0,95\mathbf{e_1} + 0,32\mathbf{e_2}$$

Les deux cas peuvent être visualisés ci-après. La configuration de $\mathbf{x_1}$ et $\mathbf{x_2}$ est choisie pour que la bissectrice entre $\mathbf{x_1}$ et $\mathbf{x_2}$ reste inchangée (axe des x).

Cas de référence :

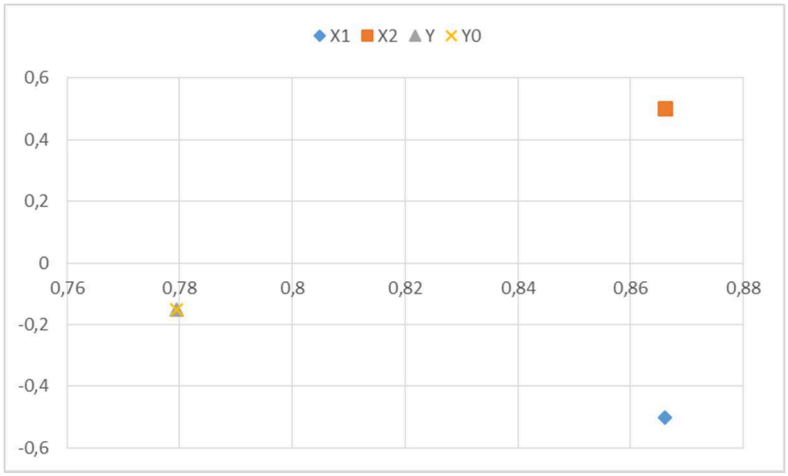


Figure 3.1.9.1: Visualisation. $b_1 = 0,6 \quad b_2 = 0,3 \quad r_{12} = 0,5$

Naturellement dans ce cas de référence $\mathbf{y_0}$ et $\mathbf{y}$ sont bien confondus.

Cas alternatif : $b_1 = 0,6 \quad b_2 = 0,3 \quad \rho = 0,8$:

Les coefficients de la régression de $\mathbf{y_0}$ sur les nouveaux $\mathbf{x_1}$ et $\mathbf{x_2}$ changent et sont désormais 0,648 et 0,174 respectivement et sont donc différents des deux valeurs de l'exemple de référence c'est-à-dire 0,6 and 0,3. Les résultats sont visualisés ci-dessous :

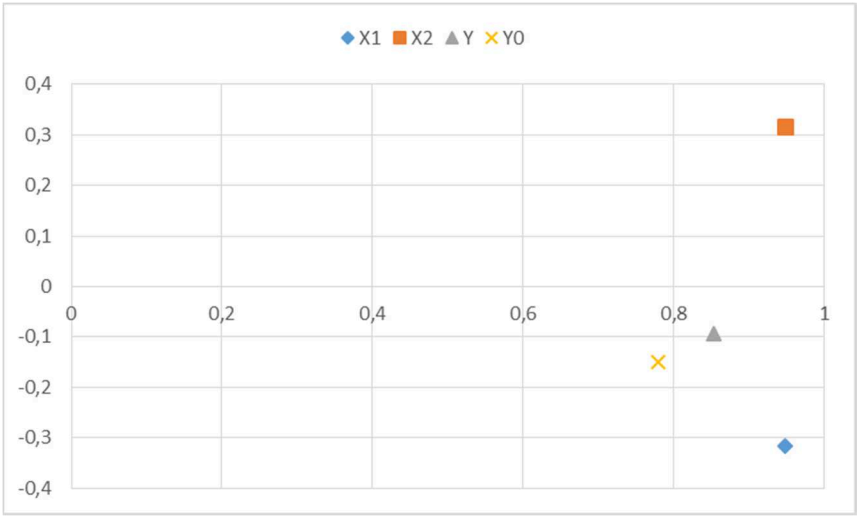


Figure 3.1.9.2. Visualisation. Comparaison de $\mathbf{y}$ et $\mathbf{y_0}$. $r_{12} = 0,8$.

Par rapport au graphe précédent cette fois $\mathbf{y}$ et $\mathbf{y_0}$ sont distincts, et $\mathbf{x_1}$ et $\mathbf{x_2}$ sont plus proches car plus fortement corrélés. Les *CAR scores* et l'écart entre *CAR scores* ont été calculés à la fois directement en utilisant les formules trigonométriques et le tableur Excel associé, mais également en utilisant le script R fourni par Zuber.

Dans le cas du scenario avec $r_{12} = 0,8$ nous pouvons calculer la différence des carrés des *CAR scores* pour $\mathbf{y}$ et $\mathbf{y_0}$.

Utilisant la formule trigonométrique ci-dessus et les calculs du tableur nous pouvons vérifier que l'écart de CAR scores calculé de cette façon pour y est identique au résultat obtenu par Zuber avec le script R soit 0,162.

En revanche la différence de *CAR scores* pour $\mathbf{y_0}$ est 0,2338269 (appelée « default value settings » par Zuber). A titre de confirmation, si cette fois nous utilisons le script R mais en entrant les paramètres correspondant à $\mathbf{y_0}$ exprimés avec les nouveaux $\mathbf{x_1}$ et $\mathbf{x_2}$ (c'est-à-dire en entrant dans le code $r_{12} = 0,8$, $b_1 = 0,648$ et $b_2 = 0,174$), en fait pour une meilleure précision en entrant les coefficients ci-après :

- Sb01= 0,647962743
- Sb02= 0,173621094

Le calcul avec R délivre exactement à nouveau la valeur de référence pour l'écart de *CAR scores* c'est-à-dire pour $\mathbf{y_0}$: 0,2338269. Ceci signifie que si y_0 est bien maintenu constant en tant que variable, même si $\mathbf{x_1}$ et $\mathbf{x_2}$ sont choisis avec une corrélation plus proche (comme 0,8) voire très proche de 1 la différence de *CAR scores* sera toujours celle du scénario de référence pour $\mathbf{y_0}$: 0,2338269

De la même façon la courbe de différence des *CAR scores* a une allure différence si au lieu des coefficients sb01 et sb02 initiaux (0,6 et 0,3) nous utilisons les nouvelles valeurs (b1=0,648 and b2=0,174) et faisons cette fois simplement varier ρ. Nous avons reproduit les résultats et joint le code R correspondant en annexe. Nous constatons que pour ρ=0,8 nous retrouvons bien la valeur de référence 0,233, ce qui correspond à la valeur attendue pour $\mathbf{y_0}$

.

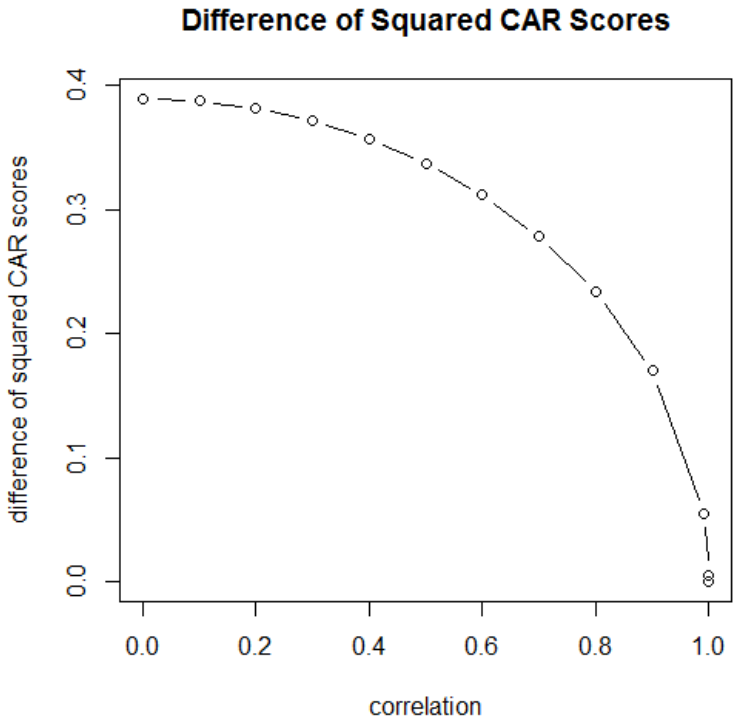


Figure 3.1.9.3 : *CAR scores* de $\mathbf{y_0}$ .Cas alternatif.

Aussi la formule analytique trigonométrique et le script R donnent bien le même résultat, cf. exemple en annexe un exemple à titre de vérification appliqué au cas de référence, c'est-à-dire : $b_1 = 0,6 \quad b_2 = 0,3 \quad r_{12} = 0,5$.

Si nous considérons maintenant un exemple de corrélation très élevée

($b_1 = 0,6 \quad b_2 = 0,3 \quad r_{12} = 0,9999$) nous pouvons visualiser les résultats suivants :

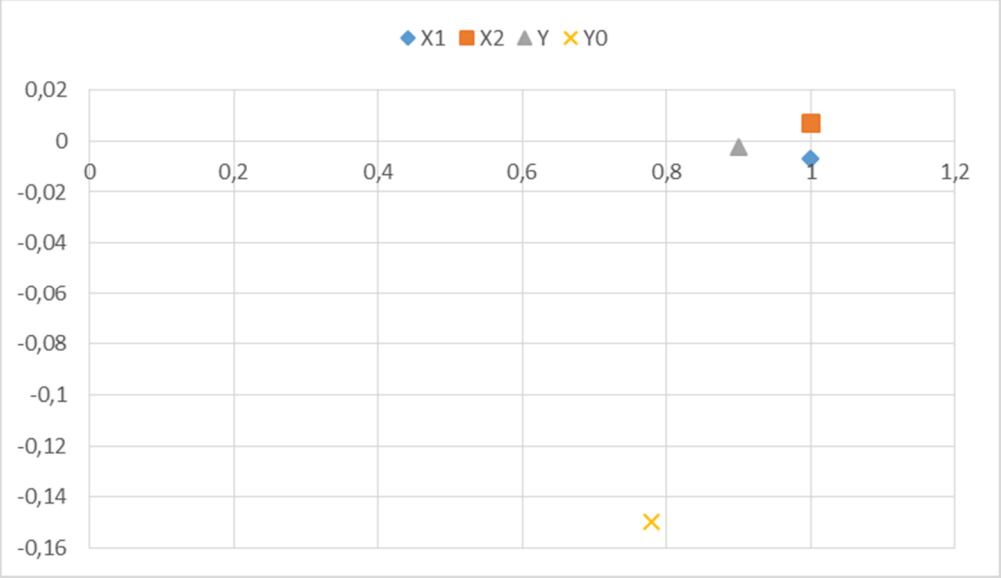


Figure 3.1.9.4 : Comparaison de $\mathbf{y}$ et $\mathbf{y_0}$. $r_{12} = 0,9999$.

$\mathbf{x_1}, \mathbf{x_2}$ et $\mathbf{y}$ tendent à être alignés et dans ce scénario $\mathbf{y}$ devient nettement différent de la référence $\mathbf{y_0}$. Comme prévu le calcul direct donne pour $\mathbf{y}$ le même écart de *CAR scores* que le script R (c.à.d. (0,0038) et pour $\mathbf{y_0}$ encore une fois l'écart de *CAR scores* est bien la valeur de référence (0,2338269).

Ces résultats proviennent du fait que la matrice des $\mathbf{Z}$ qui est utilisée pour calculer les *CAR scores* peut être la même pour des $\mathbf{X}$ de départ très différents. Notons que Zuber et Strimmer appellent cette matrice la matrice des « Mahalanobis Decorrelated Predictors ».

En reprenant les notations antérieures de la décomposition en valeur singulière : $\mathbf{X = P\Delta Q'}$ et $\mathbf{Z = PQ'}$, nous pouvons présenter les CAR scores de la façon suivante dans les cas où il y a plus de deux prédicteurs:

$$\mathbf{\Omega = QP'y}$$
$$\mathbf{\Omega} = \left(\omega_j\right)$$
$$CAR(j)^2 = \omega_j^{\,2}$$

Ceci montre que dans le cas général avec p ($p \geq 3$) prédicteurs les *CAR scores* ne dépendent que de $\mathbf{P, Q}$ et $\mathbf{Y}$, mais pas de $\mathbf{\Delta}$. Le choix d'une matrice $\mathbf{\Delta}$ correspond à une configuration donnée de colinéarité des $\mathbf{X}$ telles que nous les avons illustrées ci-dessus avec le cas des deux prédicteurs dans le plan en faisant simplement varier ρ.

Dans le cas de plus de deux dimensions il est aussi possible de générer des situations de forte corrélation (une matrice Δ avec des valeurs très hétérogènes) sans que les écarts de *CAR scores*, qui ne dépendent que de $\mathbf{PQ'}$ et de $\mathbf{y}$, ne changent.

En conclusion si $\mathbf{y}$ est bien maintenu constant (c.à.d. égal à $\mathbf{y}_0$) la différence entre *CAR scores* ne tend pas vers zéro systématiquement quand la corrélation entre les deux prédicteurs devient forte.

En fait l'utilisation de la matrice $\mathbf{Z}$ faite par Genizi ou Johnson réalloue de la variance entre les prédicteurs de départ de façon plus informative, tandis que l'approche plus simpliste des *CAR scores* ne tient plus compte suffisamment de la structure de départ des $\mathbf{X}$. La proposition de Strimmer et Zuber est donc erronée, lorsque la corrélation entre deux prédicteurs tend vers 1 la différence des carrés des *CAR scores* ne tend pas forcément vers zéro.

Commentaire 2

Strimmer et Zuber indiquent aussi que dans le cas où la configuration des prédicteurs $\mathbf{x}_1$ et $\mathbf{x}_2$ et de la variable à prédire $\mathbf{y}$ est telle que les deux variables sont positivement corrélées mais que les coefficients de régression multiple de $\mathbf{y}$ sur $\mathbf{x}_1$ et $\mathbf{x}_2$ sont de signes opposes les *CAR scores* tendent vers zéro quand la corrélation entre $\mathbf{x}_1$ et $\mathbf{x}_2$ tend vers 1 :

« In addition to the grouping property the CAR score also exhibit and important behaviour with regard to antagonistic variables. If the regression coefficients of the two variables have opposing signs and these variables are in addition positively correlated then the corresponding CAR scores decrease to zero. For example, with $(b_{std})_2 = -(b_{std})_1$ *we get*

$$\omega_1 = -\omega_2 = (b_{std})_1\sqrt{1-\rho}$$

Ce résultat annoncé par les auteurs est lui aussi erroné et est en outre en contradiction avec les autres parties de l'article qui rappellent (au 4.6 de l'article) que le coefficient de détermination R^2 est la somme des carrés des *CAR scores*, donc les « corresponding CAR scores » ne sauraient tous deux converger simultanément vers zéro.

Dans l'exemple cite ci-dessus avec $b_{st2} = -b_{st1}$ (c.à.d. $b_2 = -b_1$ avec les notations précédentes) ceci correspond en fait au cas où $\omega_1 = -\omega_2 = |y|\frac{\sqrt{2}}{2}$ et le carré des ω_1 et ω_2 est 0,5 $|\mathbf{y}^*|^2$.

Strimmer et Zuber n'ont en fait pas tenu compte du fait que quand la corrélation entre $\mathbf{x}_1$ et $\mathbf{x}_2$ tend vers 1, les coefficients de régression standardisés $(b_{std})_1$ et $(b_{std})_2$ (ou β_1 et β_2) deviennent très grands en valeur absolue et le produit par $\sqrt{1-r_{12}^{\ 2}}$ ne tend pas vers zéro.

Commentaire 3

Enfin pour des raisons semblables une troisième affirmation énoncée dans cette publication est contestable :

"This implies that antagonistic positively correlated variables will be bottom ranked. A similar effect occurs for protagonistic variables that are negatively correlated, as with $(b_{std})_1 = (b_{std})_2$ *we have* $\omega_1 = \omega_2 = (b_{std})_1\sqrt{1+\rho}$ *which decreases to zero for large negative correlation (i.e. for* $r \to -1$ *)."*

La configuration décrite dans ce dernier point, où les deux coefficients de corrélation standardisés sont égaux, avec une corrélation négative entre $\mathbf{x_1}$ et $\mathbf{x_2}$ est celle ou y est parfaitement corrélé avec la première composante principale (direction de $\mathbf{x_1} + \mathbf{x_2}$).

Dans ce cas il est vrai que $\omega_1 = \omega_2$ mais ces valeurs ne tendent pas vers zéro quand $\rho \to -1$ mais restent simplement constantes avec ω_1^2 et ω_2^2 chacune égale à $0,5\left\|\mathbf{y}^*\right\|^2$.

Commentaire 4

Dans une présentation Zuber (2010) énonce que les CAR scores vérifient la propriété de "proper inclusion" qui est identique à la propriété « inclusion » mentionnée au 3.1.2 ci-dessus : $\omega_j \neq 0 \quad si \quad \beta_j \neq 0$ (proper inclusion).

En réalité cette proposition est erronée.

Les formules ci-après montrent que la propriété de "proper inclusion" ne sera pas respectée si $\varphi \neq \psi$ et $\psi = \pi/4$.

$$\beta_1 = \frac{\sin(\varphi - \psi)}{\sin(2\varphi)}$$
$$\beta_2 = \frac{\sin(\varphi + \psi)}{\sin(2\varphi)}$$
$$\omega_1^2 = \cos^2(\psi + \frac{\pi}{4})$$
$$\omega_2^2(2) = \cos^2(\psi - \frac{\pi}{4})$$

En effet pour le prédicteur 1 la propriété de « proper inclusion » ne sera pas respectée car $\beta_1 \neq 0$ alors que $\omega_1 = 0$

.

En conclusion l'approche CAR ne paraît pas convaincante car ou bien les prédicteurs initiaux sont suffisamment décorrélés et le recours aux CAR n'est guère utile, ou bien les prédicteurs sont fortement corrélés et en ce cas les

vecteurs colonnes de $\mathbf{Z}$ ($\mathbf{PQ'}$) sont en fait trop différents des colonnes de $\mathbf{X}$ pour permettre une analyse efficace de l'importance.

Un mode de décomposition de la variance moins compliqué que *lmg-Shapley* ou la version de Johnson, et qui est aisé à formaliser dans un script R, est détaillée selon la méthode présentée ci-après, méthode dite « *weifila* ».

3.1.10 Méthode *weifila* (weighted first last)

Cette méthode est une décomposition de la variance et est une alternative moins lourde en termes de calculs que Shapley Value ou Johnson/Genizi. Dans une décomposition de la variance, il s'agit d'allouer à chaque prédicteur j une valeur w(j) telle que

$$\sum_j w(j) = R^2$$

Le calcul avec weifila est effectué comme suit :

Definition: Importance *weifila* (Weighted First Last):

Soient :

$$L = \sum_j last(j)$$

$$F = \sum_j \text{first}(j)$$

L'importance weifila est :

$$\text{W}(j) = last(j)\left(\frac{F - R^2}{F - L}\right) + first(j)\left(\frac{R^2 - L}{F - L}\right)$$

En pratique il convient de vérifier les propriétés suivantes :

1. Vérifier si $R^2 \in [L, F]$

   1.1. Si $L \le R^2 \le F$

   Calculer W(j)

$$\text{W}(j) = last(j)\left(\frac{F - R^2}{F - L}\right) + first(j)\left(\frac{R^2 - L}{F - L}\right)$$

on a bien $\sum_j \text{W}(j) = R^2$

1.2. Si $F \leq R^2 \leq L$

Calculer W(*j*)

$$\mathrm{W}(j) = last(j)\left(\frac{R^2 - F}{L - F}\right) + first(j)\left(\frac{L - R^2}{L - F}\right)$$

Le cas $R^2 \notin [L, F]$ n'est pas rencontré en pratique.

Cette méthode de calcul *weifila* est formalisée dans un script R en annexe 2. Elle donne des résultats proches de la décomposition lmg-Shapley comme le montre le calcul sur le jeu de données *swiss* :

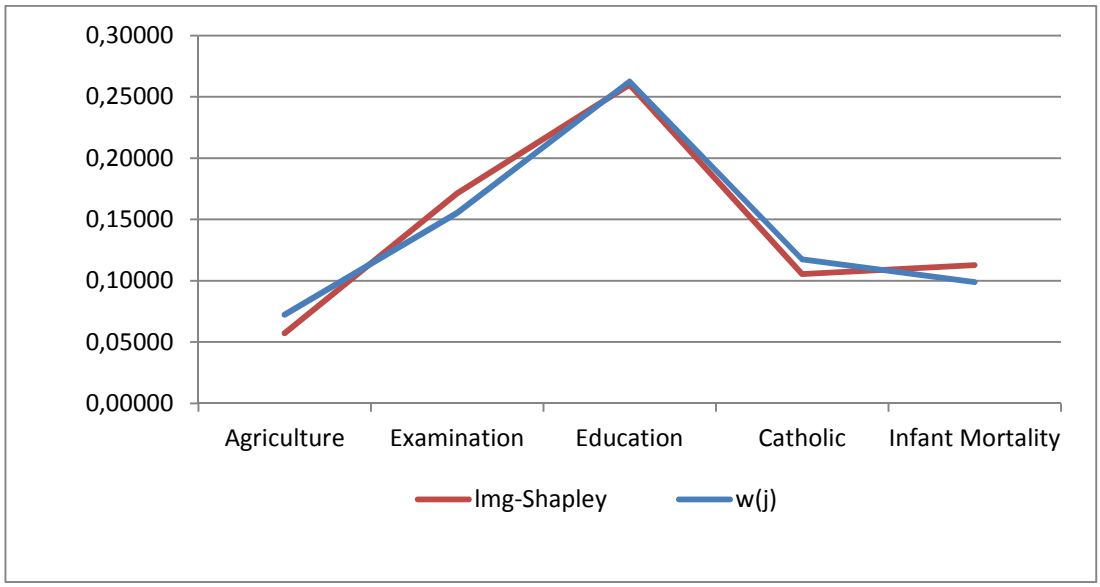


Figure 3.1.10.1. Comparaison entre *lmg-Shapley* et *weifila*. Données *swiss*.

A noter que *weifila*, combinaison linéaire des *first* et des *last* est une valeur proche de *lmg-Shapley* mais que *first* ou *last* en tant que tels, même après normalisation (c.à.d. pour que la somme soit égale au R^2 ou à 100 %) diffèrent davantage de *lmg-Shapley*, comme le montre la figure ci-après :

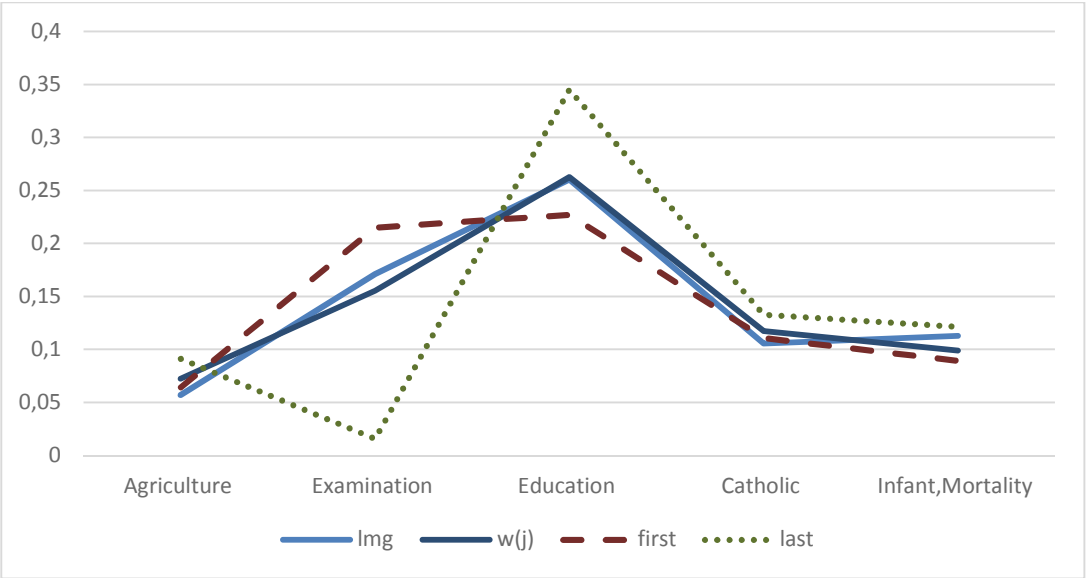


Figure 3.1.10.2. Comparaisons entre *lmg-Shapley, weifila, first* et *last*. Importances normalisées. Données *swiss*.

Voici une comparaison de la décomposition *weifila* cette fois avec *lmg-Shapley* et *Johnson* sur un autre jeu de données et 9 prédicteurs (données *comparaison)* en tableau et en graphique :

	Johnson	**lmg**	**w(j)**
Driver 1	0,060	0,063	0,056
Driver 2	0,048	0,045	0,041
Driver 3	0,026	0,027	0,033
Driver 4	0,018	0,022	0,032
Driver 5	0,026	0,023	0,027
Driver 6	0,059	0,059	0,053
Driver 7	0,061	0,060	0,055
Driver 8	0,039	0,039	0,040
Driver 9	0,043	0,042	0,044

Tableau 3.1.10.1. Décompositions de Johnson, *lmg-Shapley* et, *weifila*. Données *comparison*.

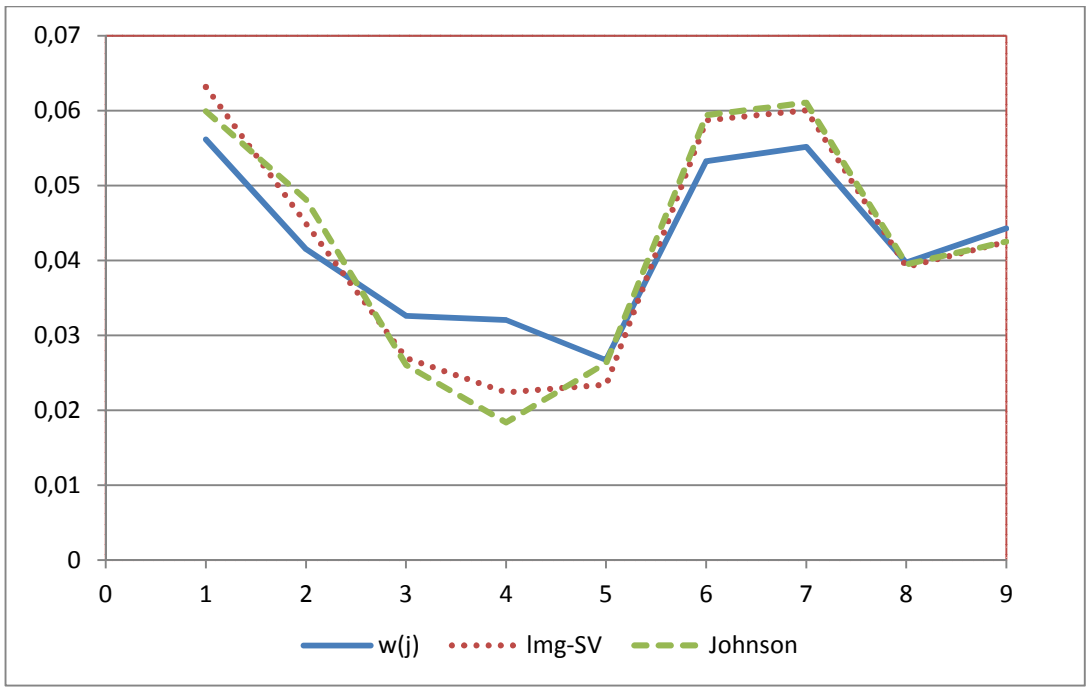


Figure 3.1.10.3. Décompositions de *Johnson, lmg-Shapley* et *weifila*. Données *comparison*.

Voici maintenant une illustration avec les données *UK Data* (1499 observations, 14 prédicteurs) :

	w(j)	lmg	Johnson	last	first
DRIVER 1	0,0193	0,0190	0,0237	0,0023	0,1551
DRIVER 2	0,0369	0,0405	0,0377	0,0048	0,2929
DRIVER 3	0,0289	0,0282	0,0276	0,0006	0,2540
DRIVER 4	0,0263	0,0243	0,0242	0,0000	0,2358
DRIVER 5	0,0258	0,0233	0,0239	0,0006	0,2266
DRIVER 6	0,0257	0,0230	0,0228	0,0000	0,2309
DRIVER 7	0,0313	0,0323	0,0311	0,0012	0,2710
DRIVER 8	0,0224	0,0205	0,0224	0,0000	0,2011
DRIVER 9	0,0287	0,0288	0,0309	0,0027	0,2361
DRIVER 10	0,0092	0,0066	0,0066	0,0003	0,0797
DRIVER 11	0,0300	0,0305	0,0300	0,0010	0,2615
DRIVER 12	0,0216	0,0177	0,0174	0,0013	0,1839
DRIVER 13	0,0313	0,0361	0,0357	0,0003	0,2783
DRIVER 14	0,0456	0,0525	0,0493	0,0125	0,3095
Σ	0,3832	0,3832	0,3832	0,0277	3,2165

Tableau 3.1.10.2. Comparaisons *weifila, lmg-Shapley, johnson, last* et *first*. Valeurs non normalisées. Données *UK Data.*

En représentation graphique focalisée sur *weifila, lmg-Shapley* et *johnson* :

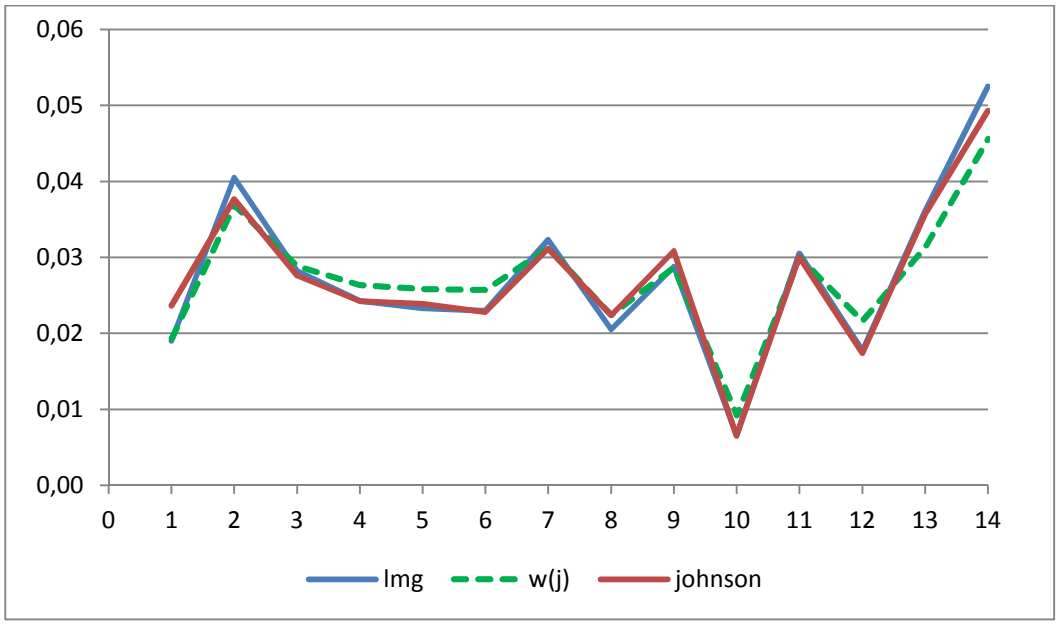


Figure 3.1.10.4. Décompositions *johnson, lmg*, *weifila*. Données *UK Data*.

Les valeurs calculées par cette dernière méthode *weifila* sont donc très proches des résultats de la méthode *lmg-Shapley* mais plus simples à calculer. En effet *lmg-Shapley* nécessite de calculer les R^2 de 2^{p-1} modèles, tandis que *weifila* nécessite seulement le calcul de $2p$ valeurs.

Vérifions également dans le cas spécifique de deux prédicteurs la cohérence des importances *weifila* avec les valeurs obtenues par *lmg-Shapley* et les Relative Weights de Johnson. Pour ceci nous pouvons utiliser à nouveau les formules détaillées précédemment :

Il vient (pour simplifier les notations comme la méthode consiste à décomposer la variance expliquée nous pouvons simplement considérer la projection de y sur l'espace des prédicteurs ce qui revient à considérer que $R^2 = 1$, ce qui simplifie les notations :

$$F = first(1) + first(2) = 1 + \cos(2\psi)\cos(2\varphi)$$

$$L = last(1) + last(2) = 1 - \cos(2\psi)\cos(2\varphi)$$

Donc :

$$F - L = 2\cos(2\varphi)\cos(2\psi)$$

Et comme :

$$w(j) = last(j)\left(\frac{F - R^2}{F - L}\right) + first(j)\left(\frac{R^2 - L}{F - L}\right)$$

Donc :

$$w(j) = \frac{first(j) + last(j)}{2}$$

en comparant aux formules de calcul de SV(j) au 2.2.3:

$$w(j) = SV(j)$$

Donc dans le cas de deux prédicteurs :

$$w(j) = SV(j) = \mathrm{RW(j)}$$

Ceci confirme la cohérence de *weifila* avec *lmg-Shapley* et *johnson,* avec donc aussi parfaite égalité dans le cas de deux prédicteurs entre les trois méthodes.

Comme indiqué auparavant, cette proximité constatée aussi avec *weifila* n'est pas en soi une preuve de validité de ces trois méthodes. En fait toutes ces méthodes consistent à calculer des importances en allouant des parts de variance expliquée qui sont situées entre les *last* et les *first*. La formule précédente proposée pour le calcul de w(j) est plus cependant simple et fournit de ce point de vue un résultat qui est tout aussi justifié, nous constatons d'ailleurs sur les exemples des données *swiss, comparison* et *UK Data*, que les w(j) sont très proches des SV et des valeurs issues de l'utilisation de la méthode de Johnson. L'étude de temps de calcul effectuée plus loin montrera que *lmg-Shapley* amène à des temps de calculs pénalisants dans les configurations correspondant à des jeux typiques des études de marchés en termes de nombre d'observations et de prédicteurs.

En complément la formule de calcul direct de Shapley Value et le fait qu'il s'agit d'une décomposition de la variance permettent de préciser un majorant de la somme des *first* et des *last* pour tous les prédicteurs :

Résultat :

Soient p prédicteurs x_j *, j=1 à j=p*

et une variable à prédire y ;

et

- *first (j) le carré de la corrélation simple entre y et* x_j
- *last(j) le carré de la corrélation semi-partielle entre y et* x_j

alors :

$$\sum_{j=1}^{j=p} first(j) + \sum_{j=1}^{j=p} \mathrm{last}(j) \leq pR^2$$

Notons que la borne supérieure est atteinte dans tous les cas s'il y a seulement deux prédicteurs, comme montré plus haut, et dans le cas où tous les prédicteurs sont identiques même si il y a plus de deux prédicteurs, en effet dans cette configuration chaque first est égal au R^2 et chaque last est nul, donc la somme est bien égale à p R^2.

La méthode de décomposition de la variance weifila est tout aussi justifiée que Shapley Value ou Johnson si l'objectif est de décomposer la variance, et cette méthode présente l'avantage d'une grande simplicité de calcul. Cette simplicité est le résultat d'une conception qui propose une valeur d'importance à chaque prédicteur comme une valeur intermédiaire entre le carré du coefficient de corrélation bi variée et le carré de corrélation semi-partielle ce qui donne assez logiquement une importance élevée si un prédicteur est corrélé avec la variable à prédire et décorrélé avec les autres prédicteurs.

Dans les études de marchés la méthode de Johnson a été substituée par certains auteurs à Shapley Value, mais cet effort paraît de peu de portée concrète vue ce qui vient d'être analysé. Le seul avantage technique réside dans le fait que le temps de calcul de la méthode de Genizi-Johnson ne croît pas sensiblement avec le nombre de prédicteurs, tandis que pour *lmg-Shapley* il est typiquement doublé avec l'ajout d'une prédicteur supplémentaire.

Certains praticiens (Conklin/Lipovetsky) ont tenté d'aller plus loin que la simple utilisation des Shapley Values et ont proposé de reconstituer les β après la phase de décomposition de la variance. La méthode développée par Conklin et Lipovetsky consiste à recalculer des β à partir des Shapley Values en résolvant p équations quadratiques, selon la formule suivante :

Pour chaque prédicteur ~~i~~ (les prédicteurs et la variable à prédire sont standardisés) :

$$\sum_{j=1}^{j=p} \rho_{ij}\beta_i\beta_j = SV(i)$$

C'est à dire en fait choisir des β tels que les Net Effects (ou les importances de Hoffman-Pratt qui comme montré supra sont équivalentes) soient égaux aux Shapley values. Nous désignerons comme $SV(\beta)$ ces coefficients ainsi calculés et de manière similaire $RW(\beta)$ les coefficients calculés en égalisant les Net Effects avec les RW de la décomposition de Johnson. Nous avons analysé sur un jeu de données les propriétés des différents coefficients par rapport aux β de la régression linéaire multiple :

Jeu Test 1

Le premier jeu de données comprenait un échantillon de 499 observations avec 9 prédicteurs et un R^2 de 0,381. Nous avons calculé pour les 9 prédicteurs les valeurs suivantes :

- $SV(j)$: Shapley Value du prédicteur j
- $RW(j)$: Relative Weights (soit Genizi-Johnson cf. 2.3.8) pour le prédicteur j
- $SV(\beta)_j$: Coefficient de la Shapley Value Regression (β tels que Net Effects = Shapley Values) pour le prédicteur j
- $RW(\beta)_j$: Coefficients de la « RW regression ». (β tels que Net Effects=Relative Weights) pour le prédicteur j

Nous avons obtenu les résultats suivants :

	OLS	SV	RW	SVBETAS	RWBETAS
Driver 1	0,19	0,06	0,06	0,12	0,11
Driver 7	0,19	0,06	0,06	0,12	0,12
Driver 6	0,17	0,06	0,06	0,11	0,11
Driver 9	0,16	0,04	0,04	0,08	0,08
Driver 2	0,10	0,04	0,05	0,09	0,10
Driver 8	0,08	0,04	0,04	0,08	0,08
Driver 5	0,03	0,02	0,03	0,05	0,06
Driver 3	-0,07	0,03	0,03	0,05	0,05
Driver 4	-0,15	0,02	0,02	0,05	0,04

Tableau 3.1.10.3

Nous constatons que les ordres sont relativement cohérents entre les méthodes que mais les rapports entre coefficients sont assez différents.

En effet les Shapley Value et les Relative Weights sont des valeurs de décomposition de la variance et ne doivent pas être assimilés à des coefficients de modèle linéaire, ce qui est malheureusement un risque lorsque des résultats sont présentés à des clients. En revanche, SV_β et RW_β sont par leur nature même des coefficients du modèle linéaire.

A partir de ces 499 observations nous avons donc calculé 5 jeux de coefficients. Nous avons utilisé le test de Vuong (Vuong (1989)) pour comparer si les 4 modèles ainsi constitués (utilisant les SV, RW, SVβ et RWβ) étaient significativement différents du modèle linéaire multiple avec les β solution des MCO.

Le test de Vuong est une application de la divergence de Kullback-Leibler.

Les étapes sont :

- Calcul de f_i et g_i pour chaque modèle $fi = \frac{e^{-\frac{(yi-yi^*)^2}{2\sigma f}}}{\sqrt{2\pi\sigma f}}$, idem pour gi

- Calcul de $l_i = \ln(\frac{f_i}{g_i})$ pour chaque observation puis de la moyenne L(f/g) et de l'écart type de cette moyenne.

- Calcul du z de L(f/g) selon une loi normale standard et test de différence avec 0.

Nous avons également effectué un F test pour chacun des 4 jeux de coefficients par rapport aux β du modèle linéaire.

Dans les deux tests nous avons trouvé que si les SV, RW ou les SVβ et RWβ issus de la résolution des équations quadratiques (Net Effects = SV ou RW) sont utilisés comme coefficients dans le modèle linéaire, les résultats sont significativement différents de celui obtenu avec le modèle standard utilisant les β.

Les deux tests sont cohérents et les SVβ et RWβ fournissent comme attendu des résultats plus proches du modèle standard avec les β de la régression.

- SV : Vuong : 4,59 F : 8,33
- RW : Vuong : 4,47 F : 7,99
- SV βs : Vuong : 2,54 F : 2,15
- RW βs : Vuong : 2,49 F : 2,06

Aussi nous avons effectué 1000 réplications bootstrap sur chaque jeu et confirmé que l'écart-type des SV(βs) de la « Shapley Value Regression » ou des « RWAβs » étaient plus faibles (environ 3 fois) que l'écart-type des coefficients de la régression linéaire.

	OLS	SV BETAS	RW BETAS
Driver 1	0,07	0,02	0,02
Driver 2	0,06	0,02	0,02
Driver 3	0,06	0,01	0,01
Driver 4	0,06	0,00	0,01
Driver 5	0,05	0,01	0,02
Driver 6	0,06	0,02	0,02
Driver 7	0,05	0,02	0,02
Driver 8	0,07	0,02	0,01
Driver 9	0,07	0,02	0,01

Tableau 3.1.10.4

Ceci montre empiriquement sur un exemple que le passage intermédiaire par des décompositions de la variance pour recalculer des βs outre sa lourdeur méthodologique se traduit par une variance réduite mais aussi par un biais significatif.

Le deuxième jeu test (*UK Data*) comprend 1499 observations et 14 prédicteurs. Le R^2 de l'OLS est 0,383.

Voici les résultats :

- dans le premier tableau les coefficients OLS et SV Betas, RW betas,
- et dans le deuxième tableau les écarts types calculés avec 1000 bootstraps.

	OLS	SVBETAS	RWBETAS
Intercept	0,00		
Driver 1	0,06	0,05	0,06
Driver 2	0,13	0,08	0,07
Driver 3	0,06	0,05	0,05
Driver 4	0,02	0,05	0,05
Driver 5	0,05	0,05	0,05
Driver 6	0,00	0,04	0,04
Driver 7	0,07	0,06	0,06
Driver 8	0,01	0,04	0,05
Driver 9	0,08	0,06	0,06
Driver 10	-0,02	0,02	0,02
Driver 11	0,06	0,06	0,06
Driver 12	-0,06	0,04	0,04
Driver 13	0,04	0,07	0,07
Driver 14	0,24	0,10	0,09

Tableau 3.1.10.5

Ecarts-Types	OLS	SV BETAS	RW BETAS

Driver 1	0,03	0,01	0,01
Driver 2	0,04	0,01	0,01
Driver 3	0,05	0,01	0,01
Driver 4	0,04	0,00	0,01
Driver 5	0,04	0,01	0,01
Driver 6	0,04	0,00	0,00
Driver 7	0,04	0,01	0,01
Driver 8	0,03	0,00	0,01
Driver 9	0,03	0,01	0,01
Driver 10	0,02	0,00	0,00
Driver 11	0,04	0,01	0,01
Driver 12	0,03	0,00	0,00
Driver 13	0,05	0,01	0,01
Driver 14	0,04	0,01	0,01

Tableau 3.1.10.6

Là aussi sur un bootstrap de 1000 tirages les écarts types observés sont réduits d'un facteur 3. Même en ne regardant que les SV β et RW β les écarts de R^2 sont significatifs par rapport au R^2 de l'OLS, donc même analyse que pour le jeu test 1.

3.1.11 Analyse de sensibilité (Sensitivity Analysis)

L'analyse de sensibilité (Sensitivity Analysis) peut être définie comme: *« The study of how uncertainty of an output model (numerical or otherwise) can be apportioned to different sources on uncertainty in the model input »* (Saltelli and al, 2004).

L'analyse de sensibilité vise à proposer un cadre d'évaluation des modèles par une approche très générale, pouvant être mise en œuvre dans des domaines variés comme l'économie, les transports, la santé, le marketing. L'analyse de sensibilité des modèles présente un lien avec la décomposition de la variance qui a été étudié par Owen (2014) et Song et al. (2014).

La première phase historique dans l'analyse de sensibilité est désignée comme l'approche locale (« local approach ») qui consiste à étudier l'impact de petites variations des inputs vis-à-vis des variations de l'output autour de valeurs données (donc « locales »).

Pour s'affranchir des limitations de l'approche locale (hypothèses de linéarité et de normalité, variations locales des valeurs d'intérêt), des approches dites « globales » (« global sensitivity analysis » ou plus loin global SA) ont été développées et sont globales en ce sens qu'elles ne reposent pas sur un jeu initial ou de référence des valeurs des inputs mais prennent en considération le modèle numérique sur l'ensemble des domaines de variation des inputs. En ce sens la « global SA » est un moyen d'étudier un modèle mathématique de façon assez générale tandis que l'approche locale se concentre sur des variations autour d'un jeu spécifique des valeurs des paramètres. Une revue des différentes méthodes de "Global Sensitivity Analysis" a été proposée par Iooss et al. 2014).

L'analyse de sensibilité permet de vérifier la pertinence et la validité de modèles (« sensivity auditing »), dans une perspective large y compris par exemple détecter les motivations éventuellement politiques d'instrumentalisation des modèles, les hypothèses sous-jacentes ou les déflations artificielles des incertitudes par exemple.

Iooss et Lemaître (2014) classent les méthodes d'analyse de sensibilité en trois grandes catégories : le screening (identification des variables les plus influents parmi de nombreuses variables), les mesures d'importances (indices quantitatifs de sensibilité) et finalement l'exploration détaillée du comportement du modèle (mesurer les effets des inputs sur l'ensemble des plages de variation). Ce dernier aspect correspond en particulier aux situations ou un modèle explicite peut être proposé et où un calcul par simulation direct est possible comme dans un modèle de chemin, ou modèle linéaire ou logistique par exemple. Les différentes approches sont analysées par Iooss et Lemaître en fonction de différents axes, principalement le nombre de modèles évalués, leur complexité et la nature de l'information analysée. Cette revue est synthétisée dans l'illustration ci-après :

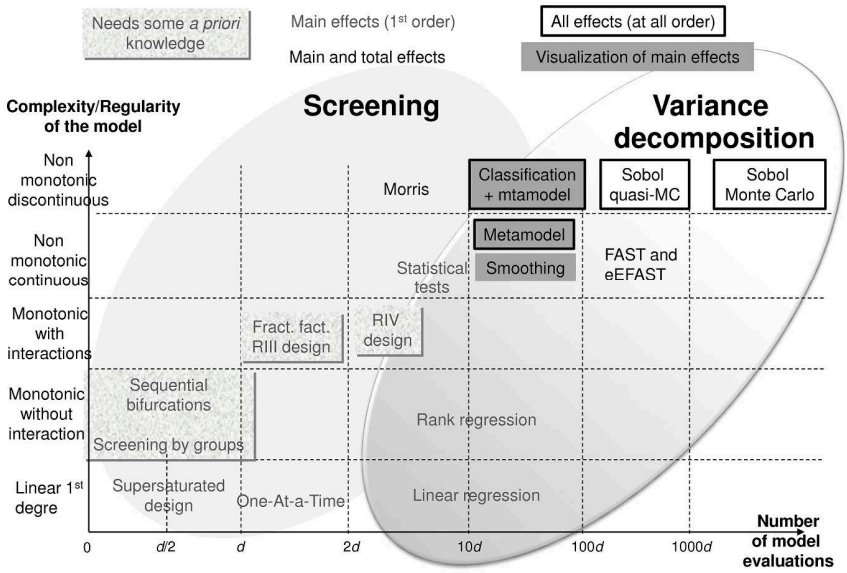


Figure 3.1.11.1 Classification des méthodes d'Analyse de Sensibilité en fonction du nombre requis d'évaluations et de la complexité des modèles. (Iooss et Lemaitre, (2014)) (avec accord des auteurs).

Saltelli (2001) a proposé une approche fondée sur la variance afin de quantifier l'importance et de déterminer des indices de sensibilité des facteurs. Il relève que ces approches supposent implicitement que ce moment d'ordre 2 est suffisant pour décrire la variabilité de l'output, mais que ceci peut ne pas être suffisant pour l'étude des queues de distribution avec des méthodes d'analyse de la variance.

Pour une présentation de l'histoire des analyses de sensibilité utilisant des méthodes fondées sur la variance, voir Saltelli en al. (2008), chapitre 4.3.

L'approche de Saltelli peut être aussi reliée aux travaux de I.M. Sobol :

Définition : Décomposition de Sobol

Soit une fonction f de carré intégrable sur Ω^k *, hyper cube unitaire de dimension k, Sobol propose la décomposition suivante :*

$$f = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + \ldots f_{12\ldots k} \qquad (3.1.11.1)$$

dans laquelle chaque terme est aussi de carré intégrable et chaque fonction f est seulement une fonction des variables référencées dans son indice. Cette décomposition comprend 2^k termes et est qualifiée de HDMR (pour High Dimensional Model Representation).

Sobol a établi que si chaque terme de la décomposition D a une moyenne nulle :

$$\int f(x_i)dx_i = 0$$

et aussi si : $$\int f_{i_1,\ldots i_s}(x_{i_1},\ldots,x_{i_s})dx_{i_k} = 0, \forall k, 1 \le k \le s$$

alors tous les termes de l'équation (2.3.11.1) sont des fonctions orthogonales c'est-à-dire :

$$\int_{\Omega_k} f_{i_1,\ldots i_s}(x_{i_1},\ldots,x_{i_s}) f_{i_1,\ldots i_t}(x_{i_1},\ldots,x_{i_t})dx = 0$$

Et Sobol (1993) a démontré qu'alors cette décomposition était unique. Notons aussi que

$$f_{0=} \int_{\Omega_k} f(x)dx$$

Les termes f ci-dessus de la décomposition de Sobol peuvent être calculés à partir des espérances conditionnelles avec la formalisation statistique des variables aléatoires :

$$f_0 = E(Y)$$
$$f_i = E(Y|X_i) - E(Y)$$
$$f_{ij} = E(Y|X_i, X_j) - f_i - f_j - E(Y)$$

En notant que :

$$V_i = V(f_i(X_i)) = V\left[E(Y|X_i)\right]$$
$$V_{ij} = V(f_{ij}(X_i, X_j)) = V\left[E(Y|X_i, X_j)\right] - V\left[E(Y|X_i)\right] - V\left[E(Y|X_j)\right]$$

Et en calculant la variance à partir de l'équation (2.3.11.1) ci-dessus il vient :

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + ...V_{12...k} \qquad (3.1.11.2)$$

Cette décomposition est appelée « ANOVA-HDMR decomposition » par Saltelli (2008).

Pour un prédicteur donné Saltelli définit deux mesures d'importance :

La première mesure est l'effet de premier ordre (first order effect) :

$$I(j) = V(E(Y|Xj))$$

cet effet traduit l'effet direct du prédicteur j. Elle est une généralisation du *betasqu* dans le cas du modèle non-linéaire.

Saltelli introduit également un indice :

$$S_j = \frac{V\left[E(Y|Xj)\right]}{V(Y)}$$

Cet indice n'est autre que le carré du rapport de corrélation $\eta^2(Y / X_j)$ bien connu.

La deuxième mesure est l'effet total (total index effect) : il s'agit cette fois d'estimer la part de variance associée à tous les prédicteurs sauf j.

Voici l'indice correspondant :

$$S_{Tj} = \frac{E\left[V(Y|X_{\sim j})\right]}{V(Y)} = 1 - \frac{V\left[E(Y|X_{\sim j})\right]}{V(Y)}$$

Dans le cas d'indépendance des prédicteurs, le first order effect est l'indice de Sobol (Sobol 1993) attribuable au sous-ensemble {j}. L'indice de Sobol a été introduit au départ pour affecter une valeur à des sous-ensembles de prédicteurs dans une partition de l'ensemble des prédicteurs en utilisant une analyse de la variance (ANOVA).

L'équation précédente liant les variances conditionnelles n'est plus respectée si les prédicteurs ne sont pas indépendants. (Saltelli et al. (2008), page 162).

Owen (2014) et Song et al (2014) ont étudié les relations entre les indices de Sobol et la Shapley Value. Owen (2014) a ainsi identifié deux fonctions de valeur (cf. 2.3.2 plus haut) qui permettent respectivement aux « first effect » et « total effect » de satisfaire les propriétés axiomatiques de la Shapley Value, et donc d'être considérées comme les Shapley Value associées à ces jeux.

Aussi il a relevé que les somme des indices de Sobol « first order », d'une part ou « total effect », d'autre part ne sont ni l'une ni l'autre égales à la variance totale. En reprenant les notations de Song et al (2014) c'est-à-dire V_i pour l'effet de premier ordre et T_i pour l'effet de deuxième ordre :

$$\sum_{i=1}^{k} V_i \leq Var(Y) \leq \sum_{i=1}^{k} T_i$$

L'égalité ci-dessus n'est atteinte qu'en cas de parfaite additivité du modèle (Owen (2014) et Song et al (2014)).

Ces considérations ont amené Owen qui s'est livré à une analyse des liens entre les indices de Sobol et la Shapley Value (Owen, 2014) à recommander la Shapley Value comme mesure d'importance dans la perspective de bonne prise en compte des interactions, ce qui illustre le lien entre décomposition de la variance et analyse de sensibilité.

Un exemple illustratif a été fourni par Song et al (2014) permettant de comparer les résultats respectifs de « direct effect », « total effect » et valeurs de Shapley.

Soit $X_K = \{X_1, X_2, X_3, X_4, X_5\}$ un ensemble de cinq variables aléatoires centrées de variance unitaire, telles que $cor(X_1, X_3) = \rho_{13}$ et $cor(X_2, X_4) = \rho_{24}$ les autres paires de prédicteurs étant indépendantes entre elles. Soit également la fonction de réponse suivante $\eta(X_K) = X_1 X_1 + X_3 X_4 + X_5$.

Alors les « first order » V_i , les « total effect » T_i et les Shapley Values Sh_i peuvent toutes être calculées :

$V_i = 0$ pour $i = 1,2,3,4$ et $V_5 = 1$

$$T_1 = T_3 = 1 - \rho_{23}^2$$

$$T_2 = T_4 = 1 - \rho_{24}^2$$

$$T_5 = 1$$

$$Sh_1 = Sh_3 = \frac{1}{2} + \frac{1}{2}\rho_{13}\rho_{13} - \frac{1}{6}(\rho_{13}^2 - \rho_{24}^2)$$
$$Sh_2 = Sh_4 = \frac{1}{2} + \frac{1}{2}\rho_{13}\rho_{13} - \frac{1}{6}(\rho_{24}^2 - \rho_{13}^2)$$

$$Sh_5 = 1$$

La variance est $V = 3 + 2\rho_{13}\rho_{24}$. Ces formules explicites permettent de montrer les écarts entre les attributions après normalisation.

Ainsi avec $\rho_{13} = 0,9$ et $\rho_{24} = 0,8$, $V_5 = 100\%$, T_5 =48%, et $Sh_5 = 23\%$.

In fine Owen recommande donc l'utilisation de la Shapley Value plutôt que des indices de Sobol.

3.1.12 Simulations

Afin d'illustrer les résultats précédents des simulations ont été réalisée avec des jeux de données générés avec les fonctions disponibles dans R.

Un jeu de données avec 1000 observations simulées a été utilisé avec seulement deux variables ($\mathbf{x_1}, \mathbf{x_2}$) pour confirmer les propriétés identifiées précédemment sur les liens entre *lmg-Shapley, johnson* et *weifila* d'une part, mais aussi sur le cas particulier de la décomposition de Fabbris avec deux prédicteurs.

La matrice de corrélation obtenue est présentée ci-après :

	$\mathbf{y}$	$\mathbf{x_1}$	$\mathbf{x_2}$
$\mathbf{y}$	1.0	0.7	0.5
$\mathbf{x_1}$	0.7	1.0	0.9
$\mathbf{x_2}$	0.5	0.9	1.0

Tableau 3.1.12.1 Matrice de corrélation. Données simulées.

Les importances non normalisées sont présentées ci-dessous :

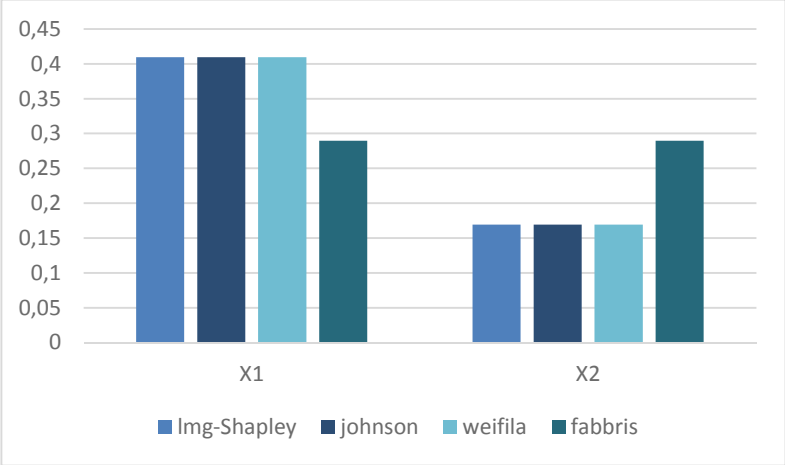


Figure 3.1.12.1. Cas de deux prédicteurs. Comparaisons *fabbris, lmg-Shapley, johnson, weifila.*

Ces résultats illustrent bien la complète égalité des importances allouées aux deux prédicteurs par *lmg-Shapley, johnson* (identique à *genizi*) et *weifila.*

Ils illustrent aussi que la décomposition de Fabbris donne une allocation égale à chacun des prédicteurs égale à la moitié du R^2 qui vaut dans cette simulation 0,58.

Ils mettent en évidente aussi que les *CAR scores* pour $\mathbf{x_1}$ et $\mathbf{x_2}$ sont très différents alors que ces deux prédicteurs sont très corrélés (0,90).

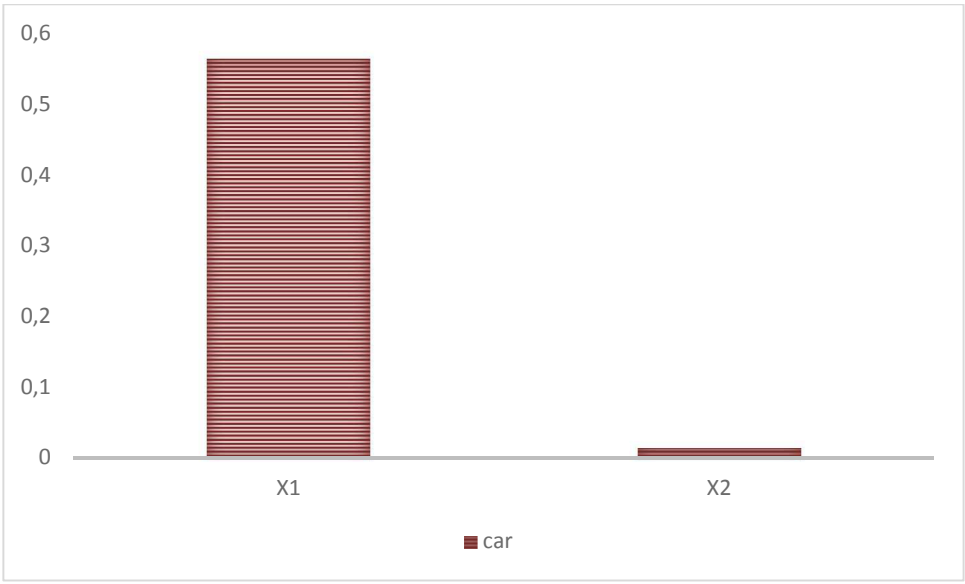


Figure 3.1.12.2. Cas de deux prédicteurs corrélés. *CAR scores*.

L'augmentation de la corrélation entre $\mathbf{x_1}$ et $\mathbf{x_2}$ dans les simulations confirme bien le maintien de cette grande différence d'allocation, ce qui illustre le résultat présenté au 3.1.9.

Un jeu de données avec 1000 observations simulées a été utilisé avec trois variables (X_1, X_2, X_3) influentes par rapport à une variable à prédire Y et trois autres variables (X_4, X_5, X_6) très décorrélées à la fois de la variable à prédire et des trois premiers prédicteurs.

Aussi une autre variable (X1Eps) a été générée avec une forte corrélation avec X_1 .

La matrice de corrélation obtenue est présentée ci-après :

	Y	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X1Eps**
Y	1,000	0,700	0,500	0,600	0,100	0,150	0,130	0,695
X1	0,700	1,000	0,650	0,650	0,100	0,050	0,100	0,995
X2	0,500	0,650	1,000	0,800	0,100	0,050	0,070	0,649
X3	0,600	0,650	0,800	1,000	0,100	0,120	0,100	0,648
X4	0,100	0,100	0,100	0,100	1,000	0,400	0,350	0,100
X5	0,150	0,050	0,050	0,120	0,400	1,000	0,280	0,048
X6	0,130	0,100	0,070	0,100	0,350	0,280	1,000	0,098
X1Eps	0,695	0,995	0,649	0,648	0,100	0,048	0,098	1,000

Tableau 3.1.12.2 Matrice de corrélation. Données simulées.

Les importances normalisées pour les sept prédicteurs sont présentées ci-dessous pour les importances *lmg-Shapley, genizi, car* et *weifila*.

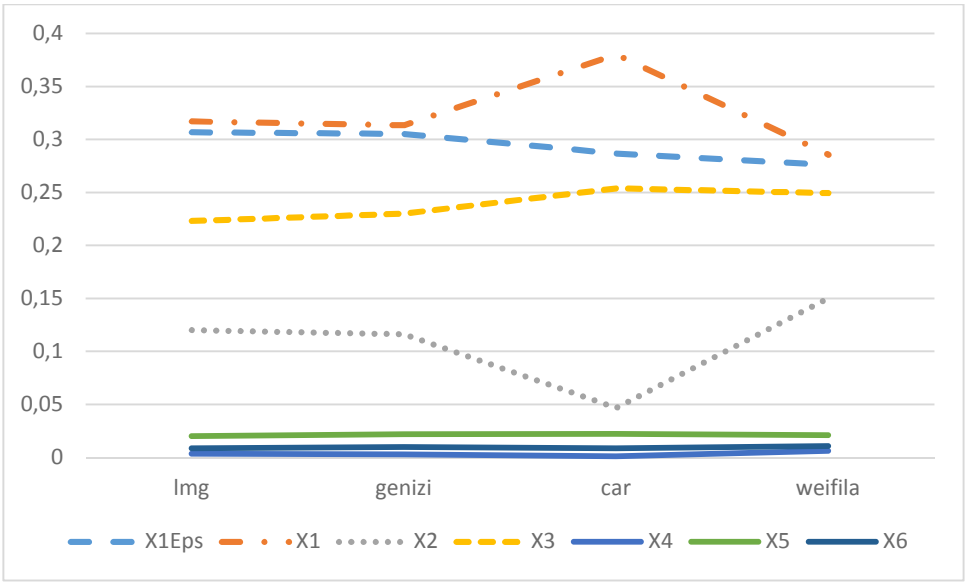


Figure 3.1.12.4. Données simulées. Résultats en % d'importance.

Dans cette simulation comme attendu seules les 4 variables influentes X_1, X_2, X_3 et X_{1Eps} reçoivent des allocations notables et les trois autres variables de bruit reçoivent bien des allocations négligeables.

Aussi il est bien observé que X_1 et X_{1Eps}, dont la corrélation est très élevée (0,99) reçoivent bien des allocations très proches avec *lmg-Shapley, genizi* et aussi *weifila*.

En revanche les *CAR scores* de X_1 et X_{1Eps} sont bien différents (0,38 % pour X_1 et 0,29% pour X_{1Eps}).

En ce qui concerne l'impact du choix de mtry dans RF-CART, la proximité entre lmg-Shapley et RF avec mtry=1 est confirmée :

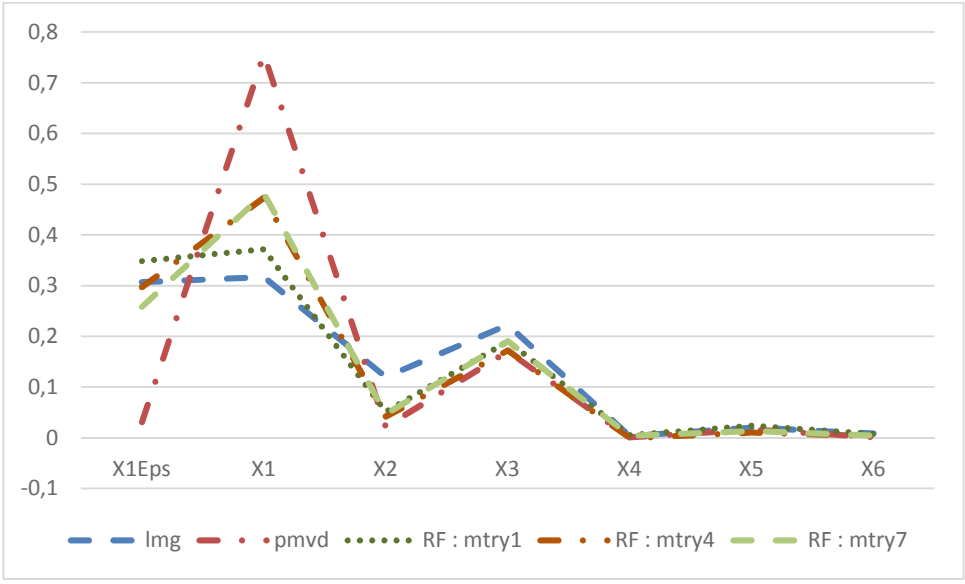


Figure 3.1.12.5. Données simulées.

3.1.13 Discussion et conclusions sur la décomposition de la variance.

La décomposition de la variance a été introduite en raison de la difficulté à appliquer pour l'analyse des leviers les recommandations des ouvrages de références, qui consistent à utiliser un nombre limité de prédicteurs et à retenir les β standardisés de l'OLS pour mesurer l'importance relative.

L'utilisation des résultats de décomposition de la variance (notamment Shapley Value et Relative Weight Analysis) a été promue dans l'activité des études de marché en raison d'une certaine stabilité dans le cas d'échantillons de taille réduite (quelques centaines d'observations) et du fait que les coefficients sont tous positifs.

Néanmoins la décomposition de la variance ne saurait en fait être un substitut à une modélisation et son utilisation inconsidérée risque de fausser l'interprétation. Nous avons montré sur des exemples de structures de variables que les proportions et l'ordre des leviers peuvent être changés par rapport l'OLS.

Aussi, l'absence de réflexion sur la réelle signification des corrélations entre prédicteurs peut conduire à accepter que plusieurs leviers isolément aient chacun un faible poids alors que si ils sont fortement corrélés ils peuvent en réalité représenter un même phénomène important comme levier d'action.

Plus fondamentalement il y a une contradiction intéressante entre vouloir d'un côté des alphas de Cronbach élevés pour conforter une cohérence entre les mesures et de l'autre côté regretter la colinéarité et ne pas tenter de l'interpréter alors que les variables mesurées sont choisies par le praticien qui formule les questions. Aussi si deux prédicteurs sont fortement corrélés l'analyse devrait plutôt porter sur une sélection éventuelle de variables ou la construction d'une variable combinée que de considérer que la réponse est de partager leur valeur d'importance comme par exemple en divisant des parts de variance.

Dans le cas des études de marchés poser plusieurs questions connexes sur un facteur important pourrait être interprété de façon erronée en considérant via l'allocation de variance qu'aucun des prédicteurs concernés n'est vraiment important alors que le facteur implicite mesuré par plusieurs réponses l'est réellement.

Nous avons montré que dans le cas de deux prédicteurs, *lmg-Shapley* et les poids de Johnson (RWA) étaient identiques, et dans le cas de plus de deux prédicteurs il a été montré sur des configurations bien choisies de décomposition par orthogonalisation que cette approche permettait d'obtenir des cas ou le *last* et le *first* d'un prédicteur donné parmi les prédicteurs de départ était atteint lors de cette décomposition par vecteurs des orthogonaux particuliers. Enfin les jeux de données utilisés ont confirmé la forte proximité des résultats entre ces deux méthodes.

Ceci a amené à proposer un mode de calcul alternatif à la Shapley Value et à la décomposition de Johnson permettant aussi d'attribuer des parts de variance expliquée à chaque prédicteur. Cette méthode de calcul (*weifila* pour *weighted first last*) est plus simple que Shapley Value ou Johnson, faisant directement appel aux corrélations bivariées et semi partielles entre les prédicteurs et la variable à prédire. Cette nouvelle approche permet sélectionner des estimateurs d'importance intermédiaires entre les *last* et les *first*, dont le résultat est identique à *lmg-Shapley* ou *johnson* pour deux prédicteurs et très proche avec plus de deux prédicteurs. En ce sens l'argument de certains auteurs (Johnson, Lebreton) selon lequel la proximité de ces deux mesures est un gage de validité intrinsèque ne nous paraît donc absolument pas fondé car ces méthodes en réalité réalisent toutes trois les mêmes choses en arbitrant entre les *last* et les *first*.

En conclusion la décomposition de la variance peut être envisagée comme un outil exploratoire mais il est déconseillé de l'utiliser comme estimateur de l'importance individuelle des prédicteurs comme si elle permettait de simuler un impact relatif, c'est-à-dire l'impact sur la variable à prédire d'un changement donné sur les valeurs d'un prédicteur (par exemple une augmentation de notation d'un point en moyenne sur un attribut dans une enquête de satisfaction). En effet comme indiqué plus haut ceci nécessite un choix de modèle et de méthode de simulation. Ceci a d'ailleurs été relevé par plusieurs auteurs (Grömping, (2007), Johnson, (2000).
Il a été également relevé dans cette recherche que c'est à tort que la méthode de Fabbris était considérée (Grömping (2015)) comme identique à la décomposition de Genizi-Johnson. Enfin plusieurs résultats concernant les CAR scores (Strimmer 2011) ont été rejetés.

A ce stade, c'est-à-dire avant de prendre en compte les avantages possibles des méthodes fondées sur des techniques plus récentes que la décomposition de la variance comme par exemple les forêts aléatoires, nous ne concluons donc pas comme Grömping (2015) à privilégier l'approche *lmg-Shapley* ou *pmvd*.

Chapitre 4 : Apports des forêts aléatoires

4.1 Introduction

Les arbres de régression et de classification (CART : Classification and Regression Trees) sont une méthode de référence en apprentissage. Elles consistent à répartir les observations en sous-ensembles de façon récursive. Pour plus de détail sur les CART voir par exemple Hastie et al. (2009). Ces outils ont été utilisés en marketing et sont aisés à comprendre et interpréter mais leur mise en œuvre sur un échantillon donné du type de ceux rencontrés dans le domaine des études de marchés présente des limitations importantes. Ainsi il est nécessaire de décider du critère de sélection d'un sous arbre pour éviter une croissance exponentielle du nombre de nœuds avec le nombre de niveaux, car sinon l'arbre devient trop grand et inutilisable car il s'ajuste trop bien aux données d'apprentissage. Les arbres peuvent présenter une instabilité au sens ou deux échantillons proches peuvent donner des résultats très différents et ne peuvent en fait être utilisés qu'avec des grands échantillons (au moins plusieurs centaines voire plusieurs milliers), ce qui dépasse souvent la taille de ceux utilisés dans l'activité d'études de marché.

Pour ces raisons les CART (Classification and Regression Trees) ont connu après une certaine popularité, un relatif oubli dans le secteur des études de marchés.

La méthode des forêts aléatoires (Random Forest) a en revanche permis de revisiter l'utilisation des arbres de régression dans le cas des études de marchés, après leur utilisation en bio statistique. Les forêts aléatoires permettent une quantification de l'importance relative (Breiman, (2001) ; Ishwaran (2007) ; Strobl et al, (2007) ; Grömping, (2009)). Dans le secteur des études de marchés cette méthode a notamment été utilisée dès 2007 par la société américaine Decision Analyst.

Nous allons d'abord présenter cette méthode puis aborder trois aspects, la comparaison entre les résultats obtenus avec les forêts aléatoires et les décompositions de la variance étudiées au chapitre 2, l'apport de ces méthodes pour la prise en compte des non-linéarités et finalement les possibilités de sélection de variables.

4.2 Présentation des forêts aléatoires (Random Forest)

Les forêts aléatoires peuvent être utilisées pour la classification ou la régression et ont été introduites par Breiman (2001). Une description très pédagogique de cette méthode a été faite par Tufféry (2015) dont nous rappellerons ci-après les principales propriétés.

La méthode des forêts aléatoires consiste à tirer des échantillons bootstrap, et à construire sur chaque échantillon du bootstrap un modèle d'une famille donnée (CART) et finalement à agréger ces modèles. Dans le cas de la régression cette agrégation se fait par la moyenne des prédictions des CART. Dans le cas du classement la méthode

retenue est fréquemment le vote. La méthode des forêts aléatoires est un perfectionnement de la méthode du « bagging » (bootstrap aggregating) dans la mesure où chaque modèle ou chaque scission des arbres, au lieu d'être effectuée avec l'ensemble des p prédicteurs, sont faits à partir d'un sous-ensemble de q prédicteurs $(q \leq p)$ tirés aléatoirement q étant constant. Quand q est strictement inférieur à p, il s'agit précisément de forêts aléatoire et non plus de bagging.

La méthode incorpore donc une randomisation tant des individus, par le bootstrap, que des variables, par la randomisation des q variables intervenant dans la construction des arbres.

Ces randomisations permettent de diminuer la variance. En effet la construction des B arbres avec le bootstrap génère des arbres identiquement distribués dont la moyenne a une espérance qui est la même que celle de ces B arbres. Mais ces arbres ne sont pas indépendants. Il est bien connu que si B variables aléatoires dans **R** sont indépendantes et identiquement distribuées et chacune de variance σ^2, la moyenne de ces B variables a une variance qui est σ^2 / B.

En revanche si ces variables ne sont pas indépendantes mais ont par exemple un coefficient de corrélation mutuel ρ alors la variance de la moyenne vaut :

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Ceci montre que si des variables sont très corrélées la variance ne sera pas fortement réduite par la croissance de B. De ce point de vue, le fait de sélectionner aléatoirement des sous-ensembles de prédicteurs permet d'éviter que les mêmes variables importantes soient sélectionnées à chaque étape de la construction des arbres et de ce fait réduit la corrélation entre les modèles produits ce qui permet donc une réduction de la variance de la moyenne des résultats des modèles.

La corrélation entre les modèles diminue quand q diminue et Tufféry (2015) relève que cette corrélation peut être proche de 0,05 ou même moins si des arbres avec q proche de 1 sont agrégés.

La recherche du bon compromis entre le biais et la variance a été étudiée et Hastie et al (2009) ont proposé que q soit égal à la partie entière de $p/3$ dans les modèles de régression et que q soit la partie entière de $\sqrt{p}$ dans les modèles de classification. Tufféry (2015) relève que des petites valeurs de q donnent souvent de bons résultats sauf si le nombre de prédicteurs discriminants est faible par rapport au nombre de prédicteurs non discriminants qui de fait représentent du « bruit ». Il indique également que *« les forêts aléatoires commencent à être très efficaces lorsque la probabilité de sélectionner un prédicteur discriminant est supérieure à 0,5 »*.

Dans la situation où il y a k variables « influentes » parmi les p prédicteurs, en choisissant q prédicteurs parmi les p à chaque étape le nombre de variables « influentes » retenues à chaque pas de construction de l'arbre suit la loi suivante :

$$P(X = x) = \frac{\binom{k}{x}\binom{p-k}{q-x}}{\binom{p}{q}}$$

La probabilité d'avoir au moins une variable influente à chaque pas de sélection de l'arbre est donc la somme suivante : $\sum_{x=1}^{x=q} P(X = x)$. (*)

Dans le cas où $q = 1$, qui sera étudié plus loin, si l'on prend en considération le critère mentionné par Tufféry (2015), pour avoir une probabilité d'au moins 50 % d'avoir à chaque pas une variable influente, il faut que le nombre de variables influentes soit au moins égal à celui des variables non influentes, ce qui est normalement le cas dans les applications d'études de marché.

La méthode des forêts aléatoires peut être formalisée de la façon suivante :
Soient une variable d'intérêt Y à valeurs dans **R** et un vecteur aléatoire $X = (X_1, ..., X_p)$ un échantillon d'apprentissage $D_n = \{(X_1, Y_1), ...(X_n, Y_n)\}$ de n vecteurs aléatoires indépendants de même loi que (X, Y) avec $X_i = (X_{i1}, ...X_{ip})$. Etant donné une fonction de régression $f(x) = E[Y/X = x]$ et $\hat{f}$ un estimateur de f , et soit $\overline{D}$ un échantillon d'observations.
L'erreur empirique de l'estimateur est calculée à partir de l'échantillon est $\overline{D}$:

$$\hat{R}(\hat{f}, \overline{D}) = \frac{1}{\left|\overline{D}\right|} \sum_{i:(X_i, Y_i) \in \overline{D}} (Y_i - \hat{f}(X_i))^2$$

Une forêt aléatoire consiste en l'agrégation d'un ensemble d'arbres aléatoires (ntree arbres) fondés sur des partitions dyadiques et récursives de l'espace des observations. A partir d'échantillons bootstrap $D_n^1, ...D_n^{n_{tree}}$ des données d'apprentissage, on engendre un ensemble appelé une «forêt » d'arbres et donc d'estimateurs $\hat{f}_1, ..., \hat{f}_{n_{tree}}$, et l'estimateur final est la moyenne des estimateurs ainsi générés.

(*) : *Cette quantité peut donc être calculée utilisant par exemple R* `>sum(dhyper(1 :q, k, p, q)`.

Ainsi qu'expliqué précédemment le principe des forêts aléatoires est d'une part de randomiser les observations par un tirage bootstrap, et aussi avec chaque échantillon de randomiser les variables en n'utilisant qu'une partie des prédicteurs, sélectionnés aléatoirement à chaque pas de construction de l'arbre, pour procéder à la séparation dyadique des observations. C'est ainsi qu'au final sera construite une forêt aléatoire.

Breiman (2001) a également proposé une mesure d'importance par permutation. Considérons la collection des ensembles « out-of-the-bag » (OOB) contenant pour chaque bootstrap les observations non retenues: $\left\{\overline{D_n^t} = D_n \setminus D_n^t, t = 1,.....n_{tree} \right\}$.

Ces ensembles permettent de définir des ensembles OOB permutés $\left\{\overline{D_n^{tj}} = D_n \setminus D_n^t, t = 1,.....n_{tree} \right\}$ en permutant aléatoirement les valeurs de la $j-ème$ variable dans les échantillons OOB. La mesure d'importance par permutation (permutation MSE) est définie par :

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[R(f_t, \overline{D}_n^{tj}) - R(f_t, \overline{D}_n^t) \right]$$

Le package R RF-CART est fondé sur les étapes suivantes :

- mettre en œuvre un bootstrap avec remplacement de l'échantillon de départ.
- générer pour chaque tirage du bootstrap un arbre avec une sélection aléatoire de prédicteurs (mtry prédicteurs est paramétrable) parmi les p prédicteurs de départ.

En moyenne 36,8 %, soit $(1/e)$, des observations sont « out of the bag » notées « OOB » pour chaque arbre.

Pour l'observation i, la prédiction par la méthode des forêts aléatoires sera $\bar{\hat{y}}_{iOOB} = RF(i)$ définie comme la moyenne des prédictions sur l'ensemble des arbres pour lesquels l'observation i est OOB.

Pour un arbre donné t on peut calculer sur l'échantillon OOB_t l'erreur de prédiction de la façon suivante :

$$errOOB_t = \frac{1}{n_{OOB,t}} \sum_{i \in OOB,t}^{n} (y_i - \hat{y}_{i,t})^2$$

où $\hat{y}_{i,t}$ est la prédiction de l'observation i pour l'arbre t.

Ensuite pour un prédicteur donné X_j en permutant aléatoirement les valeurs de X_j dans l'échantillon OOB il est possible de générer un échantillon d'observations modifié noté $\overline{OOB}_t^j$ et de calculer l'erreur de prédiction $err\overline{OOB}_t^j$ sur ce nouvel échantillon.

Ceci permet de calculer l'importance par permutation de la variable X_j :

$$VI(X_i) = \frac{1}{ntree} \sum_t (err\overline{OOB}_t^j - errOOB_t)$$

Cette mesure d'importance est aussi appelée « permutation importance ». Des études théoriques de l'importance par permutations ont été conduites par Strob et al. (2007) et plus récemment par Gregorutti et al. (2013 ; 2015). Les travaux récents de Gregorutti, Michel et Saint-Pierre quant à l'étude théorique de l'importance des variables par permutation dans les forêts aléatoires pour des modèles additifs de régression ainsi que l'extension de ces mesures à des blocs de variables laissent entrevoir des applications possibles à l'avenir pour prendre en compte de façon efficace la colinéarité dans les études de marché. Notons également que les mesures d'importance par permutations dans les forêts aléatoires peuvent être vues comme des estimateurs des indices de Sobol présentés au 3.1.11.

Remarquons que cette importance peut être légèrement négative (Genuer, Poggi, Tuleau-Malot. (2012)). Cette possibilité présente un intérêt dans la perspective de la quantification de l'importance. Notons également que cette possibilité de rencontrer des valeurs de « réduction » de la MSE négative n'est pas mentionnée dans les travaux de Grömping sur le sujet (Grömping (2009)) en ce qui concerne la moyenne sur l'ensemble des arbres ntree, même si elle mentionne que pour un arbre donné cette différence peut être négative. Rappelons que dans les travaux antérieurs sur l'importance des prédicteurs la solution d'allocation de la variance par la décomposition de Pratt par exemple a pu être écartée par certains auteurs pour cette simple raison.

Il se trouve que dans ses travaux Grömping ne rencontre pas la situation d'une réduction de MSE. Nous l'avons néanmoins rencontrée au cours de notre recherche (cf. plus loin dans ce chapitre les analyses prostate data, dans le cas de la variable xage) et avons aussi constaté que cette question était mentionnée dans le blog statexchange.
Aussi cette mesure d'importance n'est pas directement comparable aux valeurs obtenues en allouant des parts de variance expliquée ou décomposant cette quantité. C'est pourquoi, dans les comparaisons, les valeurs d'importance relative sont normalisées pour que leur somme soit égale à 1 ou à 100 %. Grömping (2009 et 2015) relève à ce propos que la normalisation et sommation à 100 % des métriques n'est pas recommandée pour des analyses de données, mais qu'elle est pratique pour comparer des résultats obtenus par plusieurs méthodes de conceptions différentes.

4.3 Application des forêts aléatoires et comparaisons

L'utilisation des forêts aléatoires dans le secteur des études de marché reste à ce jour quasi-inexistante, mais plusieurs résultats présentés par Grömping (2009 et 2015) et Genuer, Poggi et Tuleau-Malot (2008 ; 2012) traitent de potentiellement utiles dans ce domaine.

La société américaine d'études des marché Decision Analysts a présenté l'utilisation de forêts aléatoires en comparaison à l'OLS d'une part et aussi en comparaison avec une autre méthode de quantification de l'importance : la méthode MaxDiff ((Marley(2005) et Colias (2007)).
La méthode MaxDiff consiste à demander à des répondants de sélectionner parmi des attributs (prédicteurs) le plus important et aussi le moins important et de calculer la différence de ces notes.

Cette méthode est donc fondée sur une importance « déclarée » (stated importance) et non sur une importance « calculée » (derived importance). Or les répondants tendent à déclarer que beaucoup d'attributs sont importants, à l'extrême même que tout est important c'est pourquoi il est en général recommandé pour évaluer l'importance des prédicteurs d'utiliser des modèles analytiques et non de se fonder sur l'importance déclarée.

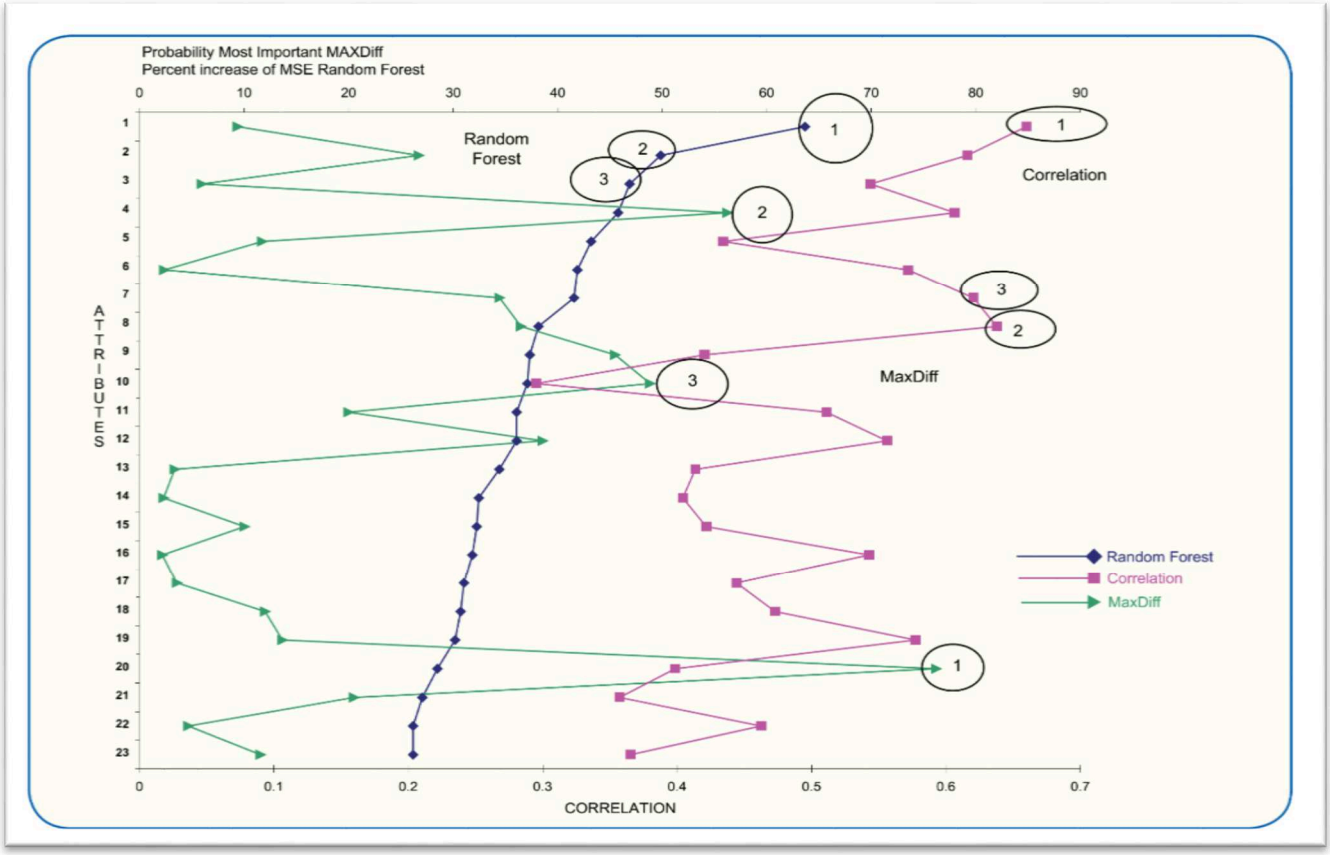


Figure 4.3.1: Comparaisons de mesures d'importance. Colias (2007).(avec accord de l'auteur).

Ce document conclut que les valeurs MaxDiff ne sont pas en fait exploitables et que les corrélations bivariées et les forêts aléatoires identifient de façon assez cohérentes les 2 ou 3 leviers les plus importants mais diffèrent ensuite dans les rangements d'importance. De manière plus surprenante John Colias conclut que les forêts aléatoires paraissent distinguer l'importance relative entre des variables corrélées.

Grömping (2009) a aussi étudié la comparaison entre régression linéaire et forêts aléatoires. Dans le jeu de données utilisé par Grömping, il y a une paire de prédicteurs particulièrement corrélés (Examination et Education avec une corrélation de 0,79) et *lmg-Shapley* égalise les valeurs d'importance par comparaison à *pmvd* ce qui est cohérent avec les résultats précédents.

Grömping relève également sur les exemples qu'elle étudie qu'avec mtry=1, *lmg* et RF-CART sont proches et conduisent tous deux à une certaine égalisation des contributions de prédicteurs corrélés entre eux mais d'influence modeste sur la variable à prédire. Aussi Grömping relève que dans le cas de mtry plus élevés CART-CI tend à se

rapprocher de PMVD tandis que les allocations avec RF-CART tendent à être plus stables malgré les évolutions de mtry. Elle souligne également que les intervalles de confiance calculés par bootstrap (et disponibles dans le package relaimpo qu'elle a développé) sont importants pour la mesure *pmvd*, qui est très variable (ce résultat de variabilité avait été déjà présenté dans sa publication de 2007). Elle relève également que RF-CART apparaît assez stable quand mtry évolue. Nous verrons plus loin que cette constatation mérite d'être nuancée et que dans certaines configurations l'impact de mtry avec RF-CART peut assez prononcé s'il y a un peu plus que quatre prédicteurs comme dans les cas étudiés par Grömping et avec plus de colinéarité entre certains prédicteurs

Grömping est revenue en 2015 sur l'utilisation des forêts aléatoires pour la quantification de l'importance, en utilisant à nouveau le jeu de données *swiss* 182 mais cette fois sans effectuer de transformation quadratique sur les deux variables « Agriculture » et « Catholic » et sans effectuer de calculs avec les forêts aléatoires. Ce calcul et cette comparaison sont faits ci-après.

Grömping (2009) ne paraît pas tirer de conclusion très nette sur l'impact de mtry sur les résultats. Elle indique que dans ses simulations (cf. figures ci-dessus) les valeurs de mtry qui minimisent errOOB dépendent du modèle (c .a .d du choix de la structure et de la corrélation entre prédicteurs) mais que ce minimum était en général obtenu pour $mtry = 1$ ou 2. Elle cite également les recommandations de Breiman d'utiliser $mtry = p/3$ pour la regression avec forêts aléatoires et relève aussi que Genuer, Poggi et Tuleau-Malot (2008) indiquent une meilleure prédiction avec mtry=p pour les données de Friedman (1991). Rappelons que Breiman (2001) recommande l'utilisation de $mtry = \sqrt{p}$ pour la classification et $p/3$ pour les regression forests.

Cependant l'approche de Grömping étant limitée au cas de 4 prédicteurs, ne permet pas une analyse suffisamment détaillée de l'effet de mtry pour les applications typiques des études de marchés où le nombre de prédicteurs est plus élevé (de un à quelques dizaines de prédicteurs). Aussi nous avons voulu comparer les parts d'allocation selon que *lmg*, *pmvd* ou RF-CART sont utilisés avec des jeux de données plus représentatifs ce qui sera fait ci-après avec davantage de prédicteurs.

RF-CART a été appliqué à différents jeux de données et aussi à *swiss 182* dans les versions quadratisées et non quadratisées afin de prolonger les analyses de Grömping. Les résultats de ces différents exemples sont présentés ci-après sous forme graphique.

Données Prostate

Ces données comprennent 97 observations et 8 prédicteurs. Voici d'abord une vue d'ensemble des importances calculées

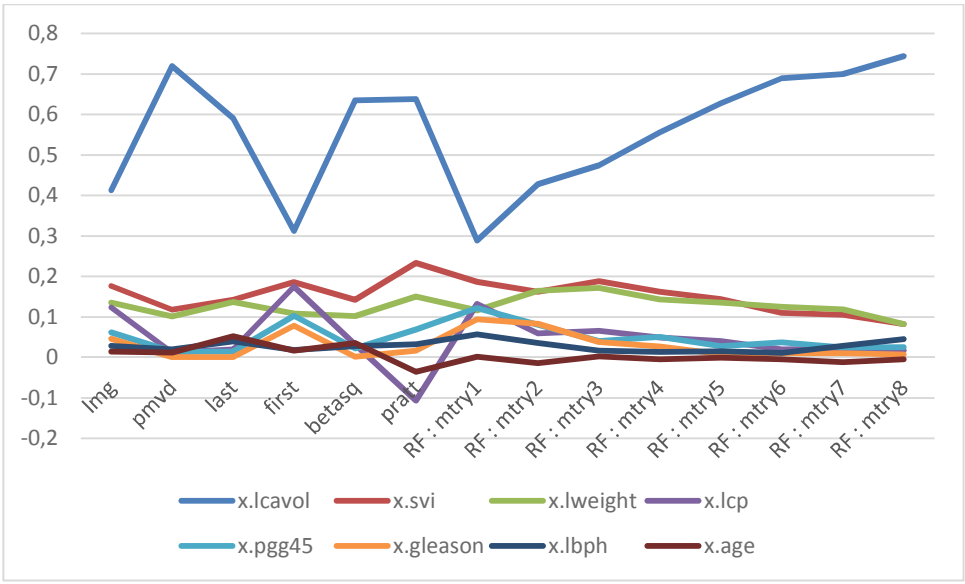


Figure 4.3.5. Comparaison des Importances Relatives. Données Prostate.

Pour la variable « âge », la permutation MSE est négative dans quelques cas. Par ailleurs comme relevé par Grömping, la proximité entre lmg et les valeurs RF mtry=1 ou 2 est là aussi constatée, et aussi la proximité entre *pmvd-betasquare* et RF-CART avec des mtry proches de p. Ces constatations sont mieux visibles ci-dessous :

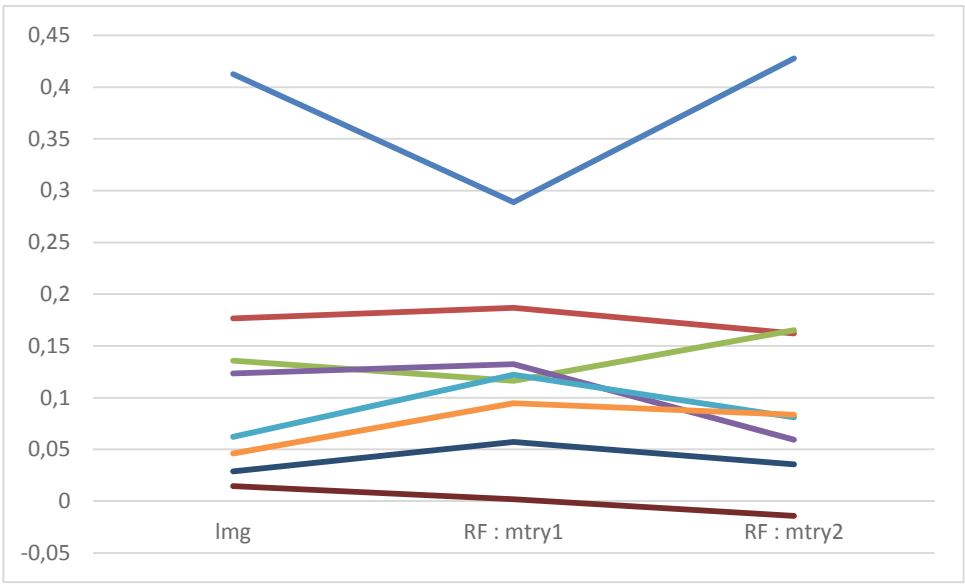


Figure 4.3.6. Comparaison lmg-mtry=1 et 2. Données Prostate/

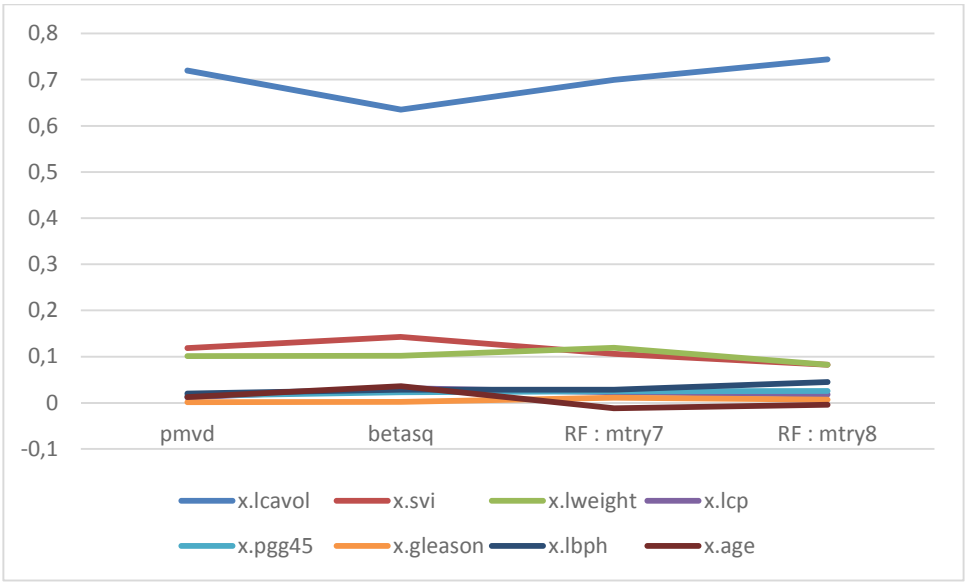


Figure 4.3.7 Comparaison *pmvd-betasqu*-mtry=7 et 8 .Données *Prostate.*

Données Credit

Les données *credit* comportent 499 observations avec 10 variables dont une variable à prédire et 9 prédicteurs.

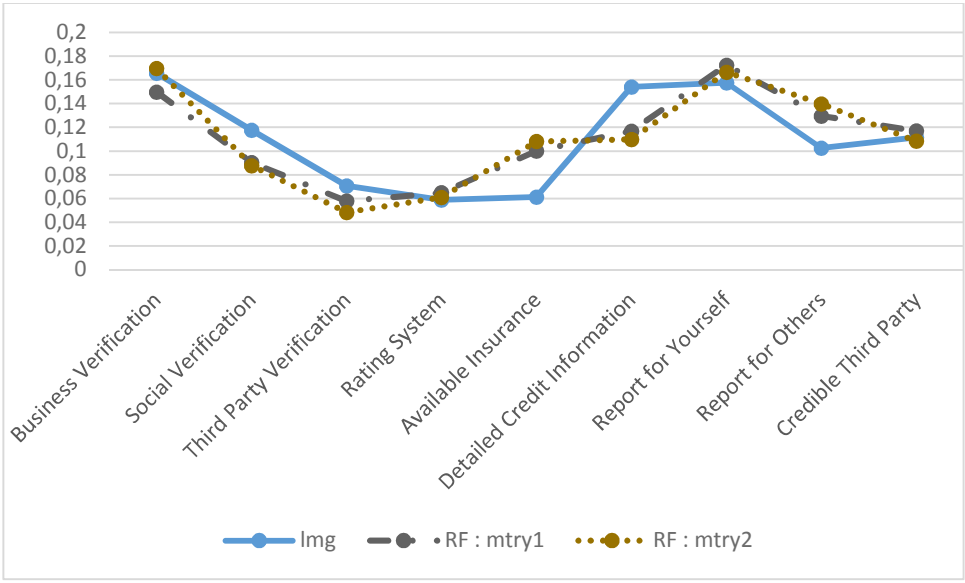


Figure 4.3.8. Comparaison lmg-mtry=1 et 2. Données Credit.

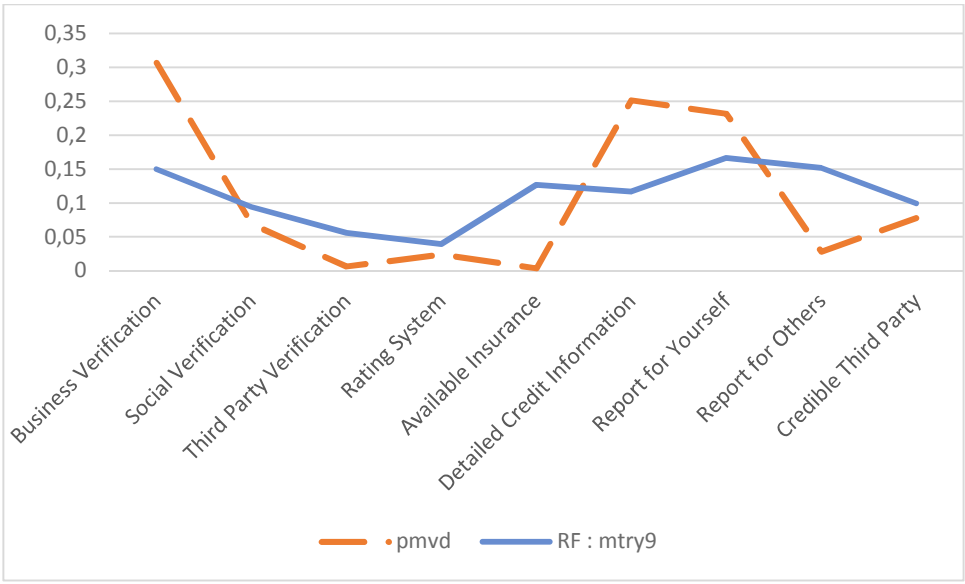


Figure 4.3.9 .Comparaison pmvd -mtry=9.Données Credit.

UK Data

Dans le cas *UK Data*, les relations entre la variable à prédire et le prédicteur sont assez linéaires et l'introduction de prédicteurs quadratisés pour les 14 prédicteurs ne conduit qu'à une augmentation réduite du R^2 : 0,3831 avec prédicteurs non quadratisés et 0,3984 avec quadratisation.

Cette situation est différente de celle rencontrée avec le jeu *swiss* 182, où l'introduction de variables quadratisées améliore plus sensiblement le R^2 du modèle. Il s'ensuit que dans le cas *UK Data lmg-Shapley* sera d'emblée proche de RF-CART avec mtry=1 en importance relative normalisée. La proximité entre *lmg-Shapley* et RF mtry1 d'une part, et *pmvd-betasquare* (et modérément de RF mtry=p) d'autre part, est de nouveau effectivement constatée et présentée dans les différentes visualisations suivantes. Les leviers ont été classés par ordre décroissant de valeur dans la décomposition *lmg-Shapley*. Différents graphes ont été édités pour faciliter la visualisation.

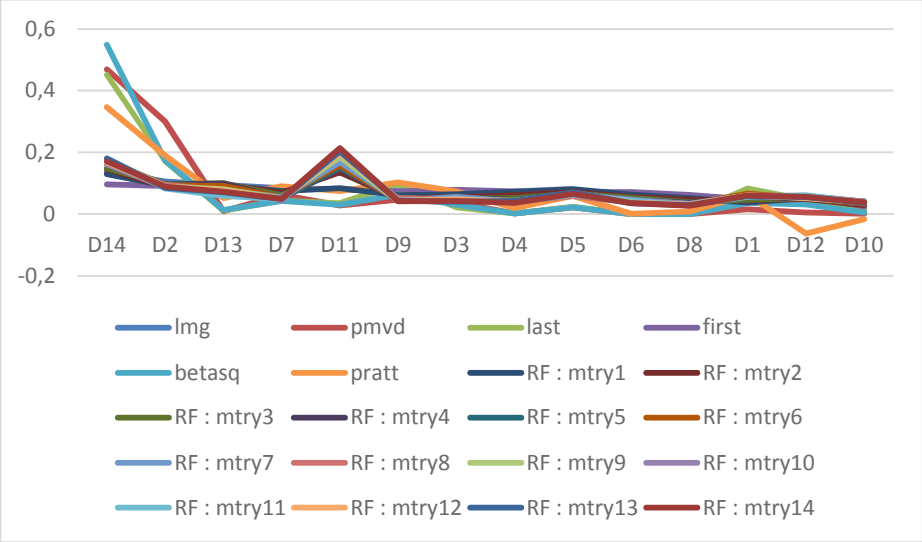


Figure 4.3.10. UK Comparaison des Importances Relatives. *UK Data*.

Comme attendu *weifila* et *lmg-Shapley* sont là encore proches et donc aussi proches de RF-CART avec mtry1.

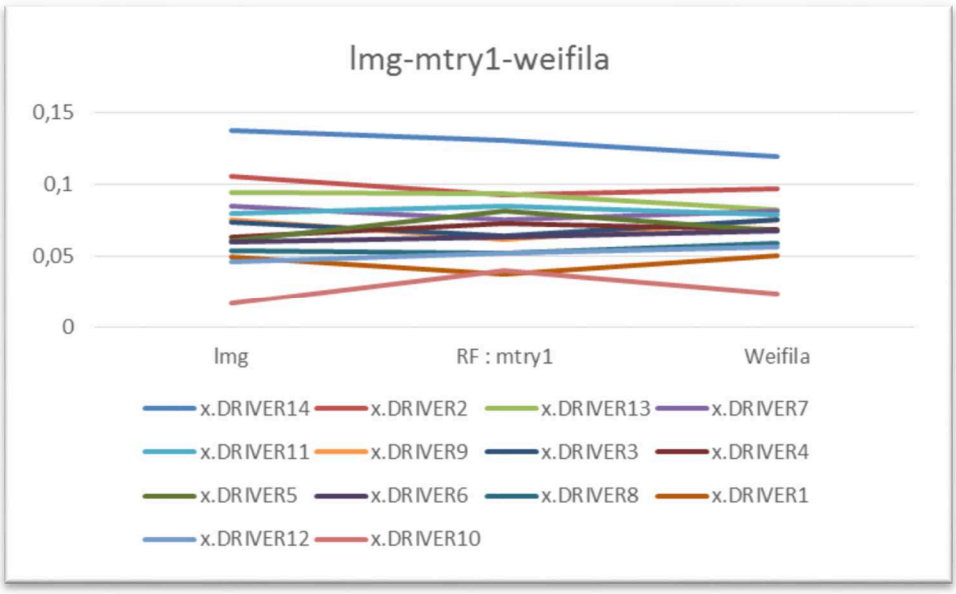


Figure 4.3.15. *lmg, weifila* et mtry=1.*UK Data*

En conclusion sur le jeu *UK Data* la cohérence entre les mesures lmg et mtry=1 est manifestée par l'analyse visuelle et par exemple par le coefficient de corrélation entre les valeurs d'importance plus élevé entre lmg et mtry = 1 (0,86) qu'entre lmg et mtry=4 (0,78) et qui décroit encore avec mtry=14 (0,59).

Inversement le lien entre les variations de mtry et betasq ou PMVD est moins visible et moins régulier. A ce sujet, alors que Grömping indique que les résultats de RF-CART sont peu sensibles au choix de mtry ceci était justifié dans les cas qu'elle a examinés avec un nombre restreint de prédicteurs (4 ou 5). En revanche avec un nombre plus élevé de prédicteurs, typiquement au moins 10, l'effet de la randomisation des prédicteurs à chaque split dans les Forêts aléatoires est plus manifeste, et la propension à l'égalisation quand mtry=1 tend à permettre aux variables ayant une moindre importance « conditionnelle » au sens utilisé par Grömping de gagner en importance et à se rapprocher de l'égalisation générée par le calcul lmg. Cet effet n'est pas aussi sensible quand le nombre de prédicteurs est trop restreint.

Voici une sélection de quelques mesures d'importance recalculées dans le cadre de cette recherche en ajoutant la décomposition *weifila* :

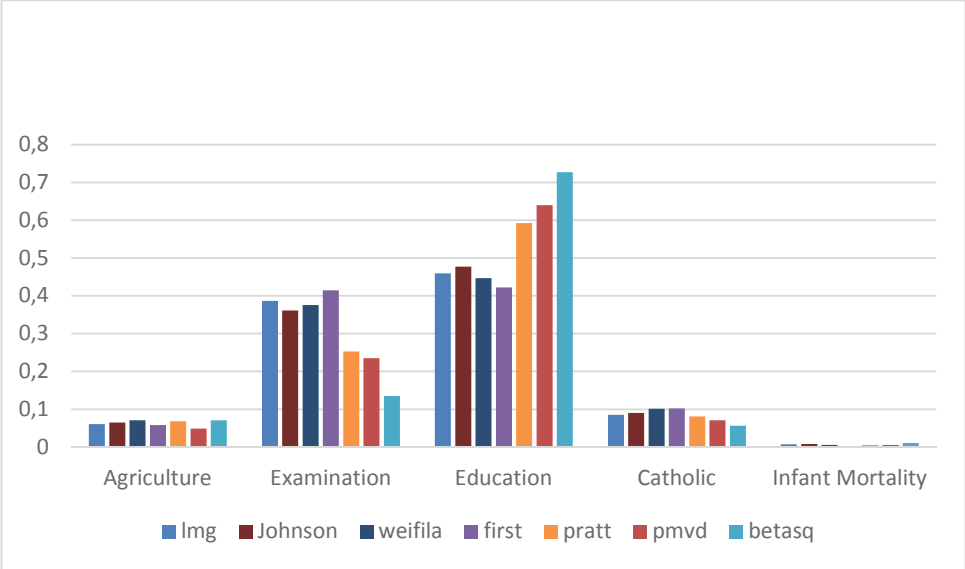


Figure 4.3.16 Comparaisons des importances. Données *swiss* 182. Calcul avec *weifil*a.

A nouveau est mise en évidence nette proximité entre *lmg-Shapley, johnson et weifila,* et une relative proximité entre *betasq, pmvd* et *pratt.*

En ce qui concerne l'influence de mtry, Genuer, Poggi et Tuleau-Malot (2008) ont relevé dans les situations standard (n>>p) et dans le cas de la régression et avec des jeux incluant de 4 à 13 variables, que la valeur par défaut n'apparaissait pas optimale et que pour des problèmes de régression standard, en ce qui concerne l'OOB error, les forêts aléatoires n'apportait pas d'amélioration par rapport au bagging sans élagage (unpruned bagging) c'est-à-dire mtry=p. Il paraît donc intéressant de réaliser une approche RF-CART sur les données de *swiss*182 sans transformation quadratique pour prolonger la synthèse présentée par Grömping en 2015, et aussi comme elle n'a analysé en 2009 l'impact de mtry que dans des cas avec 4 prédicteurs. La comparaison des résultats des obtenus avec données *swiss*182 selon que les variables sont quadratisées ou non sera effectuée ci-après, suivie par des simulations pour confirmer les résultats obtenus sur ce jeu particulier.

4.4 Forêts aléatoires et non-linéarités.

L'examen des résultats présentés par Grömping dans deux articles parus en 2009 et 2015 avec un même jeu de données analysées une fois avec et une fois sans transformation quadratique de certaines variables a montré des différences sensibles des décompositions allouées entre les prédicteurs. Dans notre recherche il a paru intéressant d'analyser les propriétés respectives et les liens éventuels entre les méthodes de décomposition de la variance étudiées au chapitre 3 et les forêts aléatoires.

4.4.1 Analyse avec les données *swiss 182*

Les mêmes données utilisées (*swiss* 182) ont été analysées différemment par Grömping : en 2009 en effectuant une transformation quadratique de deux variables (« Agriculture » et « Catholic ») et en 2015 sans effectuer cette transformation. Les visualisations ci-dessous illustrent le caractère quadratique des deux prédicteurs « Agriculture » et « Catholic » vis-à-vis de la variable à prédire « Fertility » :

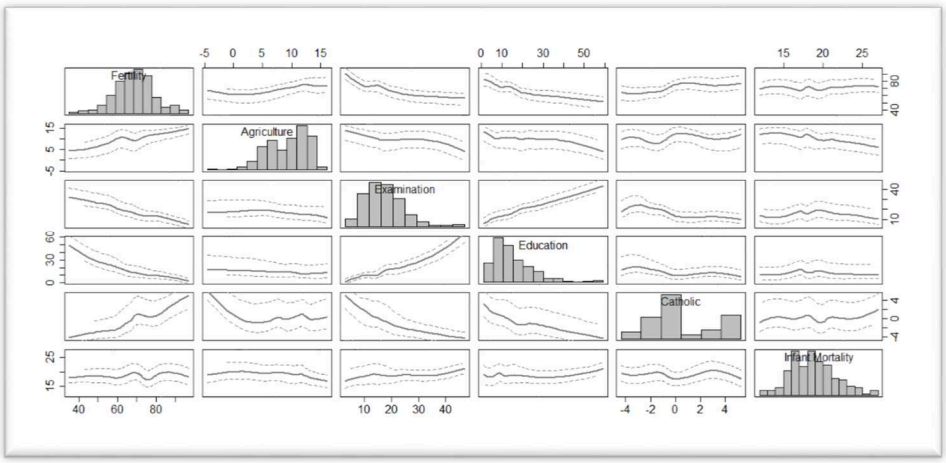


Figure 4.4.1. Données *swiss* 182. Variables non quadratisées.

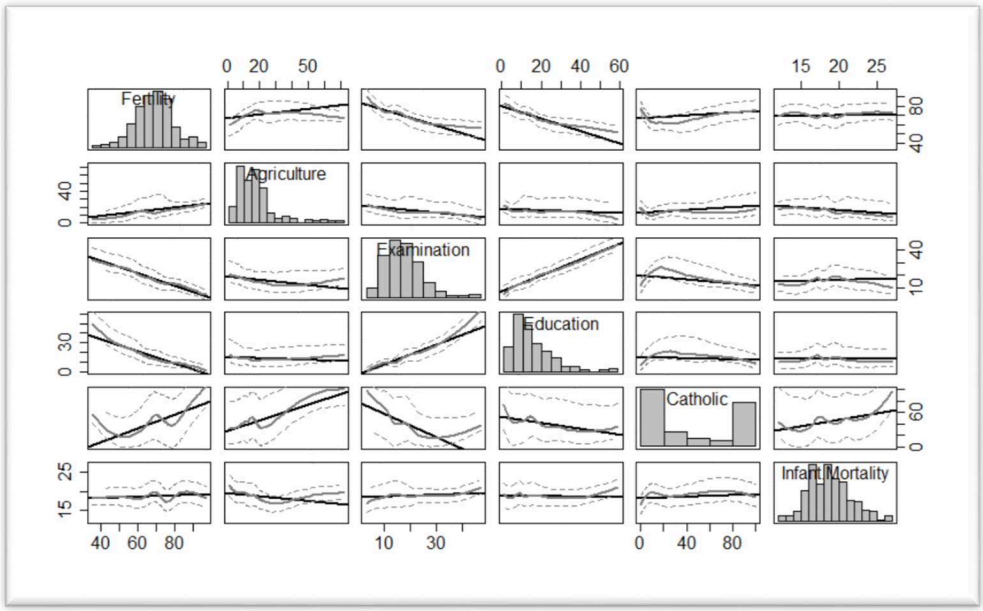


Figure 4.4.2. Données *swiss* 182. Variables « Agriculture » et « Catholic » quadratisées.

Nous avons recalculé les importances avec les forêts aléatoires en utilisant les données de 2015 (non quadratisées) pour les comparer avec les décompositions obtenues avec les variables quadratiques présentées dans la publication de 2009.

Pour reproduire les résultats de Grömping en 2009 nous avons procédé de la façon suivante : les données quadratisées sont obtenues en effectuant d'abord une régression de « Fertility » sur les 7 prédicteurs comprenant les 5 initiaux et en ajoutant les carrés des 2 variables « Agriculture » et « Catholic » puis en calculant des variables « quadratisées » « Agriculture Q » et « Catholic Q » à partir des coefficients de cette régression :

$$AgricultureQ = 0,9520149\,Agriculture - 0,0141127\,Agriculture^2$$

$$CatholicQ = -0,2158318\,Catholic + 0,0026831\,Catholic^2$$

Nous retrouvons bien le R^2 présenté par Grömping dans son article de 2009 (R^2 =61,3%) : la valeur avec les variables quadratisées calculées ci-dessus est $R^2 = 0,6127$. De la même façon en utilisant cette fois les données non quadratisées (Grömping 2015) nous retrouvons bien le même R^2 de 0,502. Et voici également les matrices de corrélations respectives :

SWISS Quadratic (2009)	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	1,000	0,426	-0,639	-0,645	0,490	0,049
Agriculture Q	0,426	1,000	-0,271	-0,187	0,035	-0,320
Examination	-0,639	-0,271	1,000	0,790	-0,577	0,061
Education	-0,645	-0,187	0,790	1,000	-0,390	-0,020
Catholic Q	0,490	0,035	-0,577	-0,390	1,000	-0,015
Infant.Mortality	0,049	-0,320	0,061	-0,020	-0,015	1,000

SWISS Non Quadratic (2015)	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	1,000	0,240	-0,639	-0,645	0,317	0,049
Agriculture	0,240	1,000	-0,199	-0,068	0,276	-0,171
Examination	-0,639	-0,199	1,000	0,790	-0,403	0,061
Education	-0,645	-0,068	0,790	1,000	-0,132	-0,020
Catholic	0,317	0,276	-0,403	-0,132	1,000	0,158
Infant.Mortality	0,049	-0,171	0,061	-0,020	0,158	1,000

Tableau 4.4.1.1 Matrices de Corrélations. Données *swiss* 182. Comparaison avant et après quadratisation.

Les coefficients de corrélations entre « Fertility » et les variables quadratisées sont dans les deux cas nettement accrus par comparaison avec les variables non quadratisées, et proches des maximums $rMax$ représentés par le coefficient de corrélation de la régression de Fertility sur seulement les deux variables Agriculture et Agriculture Q, d'une part et Catholic et Catholic Q d'autre part :

$$r(Fertility, Agriculture) = 0,240 \qquad r(Fertility, AgricultureQ) = 0,426 \qquad rMax = 0,434$$

$$r(Fertility, Catholic) = 0,317 \qquad r(Fertility, CatholicQ) = 0,490 \qquad rMax = 0,494$$

Données *swiss 182* sans effet quadratique (Grömping 2015)

Comme indiqué précédemment dans la publication de 2015, Grömping a utilisé les données *swiss* 182 sans quadratisation, alors que dans la publication de 2009 deux des variables avaient été quadratisées, « Agriculture » et « Catholic ». Le tableau suivant présente une vue d'ensemble des différentes décompositions dans le cas des variables non quadratisées.

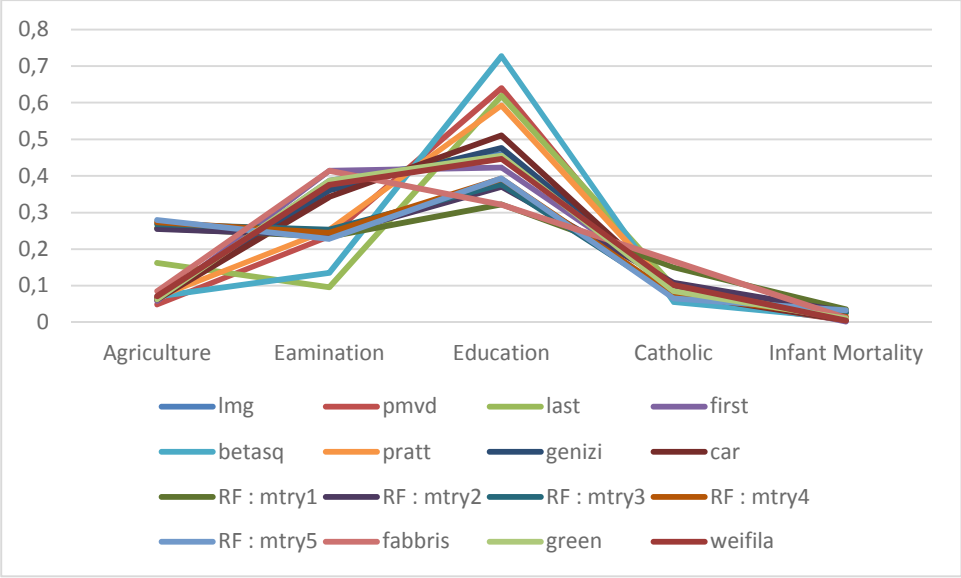


Figure 4.4.1.1. Comparaison d'importances normalisées. Données *swiss* 182 non quadratisées.

Le tableau ci-dessous permet de comparer les allocations obtenues avec *lmg-Shapley* et de les comparer à celle obtenus avec RF-CART mtry=1 et weifila.

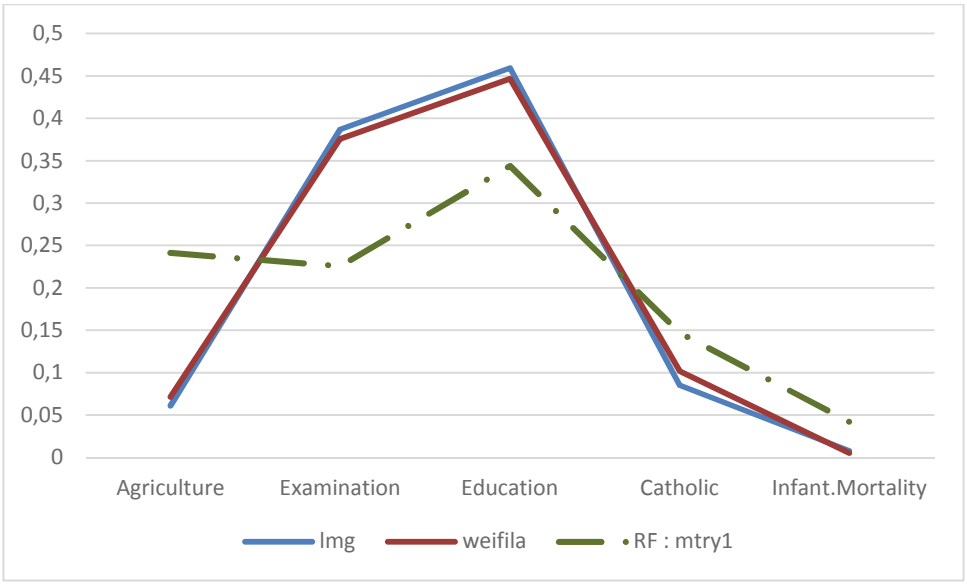


Figure 4.4.1.2. RF mtry=1 versus *lmg* et *weifila*. Données *swiss* 182 non quadratisées.

Cette première analyse montre une différence assez nette entre les allocations de *lmg-Shapley* calculées sur les données non quadratisées et RF-CART avec mtry=1 sur ces mêmes données.

Ces différences sont détaillées numériquement ci-après. Les poids d'« Agriculture » et de « Catholic » sont nettement renforcés par la quadratisation (cf. tableau ci-dessous).

lmg	**2009 (quadratisées)**	**2015 (non quadratisées)**
Agriculture	22,0%	6,1 %
Catholic	18,3 %	8,5 %

Tableau 4.4 1.1. % d'importance *lmg-Shapley*. Données *swiss* 182 quadratisées ou non.

De façon liée « Examination »et « Education » (rappelons que ces deux variables sont très corrélées) sont évaluées très différemment dans les deux cas et voient leurs poids respectifs diminuer avec la quadratisation d' « Agriculture » et « Catholic ».

lmg	**2009 (quadratisées)**	**2015 (non quadratisées)**
Examination	25,6%	38,7 %
Education	31,5 %	45,9 %

Tableau 4.4.1.2. % d'importance *lmg-Shapley*. Données *swiss* 182 quadratisées ou non.

L'effet de répartition est de nouveau bien illustré pour des variables corrélées comme le sont Examination et Education (ρ=0,79) la somme de leurs valeurs allouées avec pmvd et lmg est très proche , ce qui fait bien sur penser aussi aux propriétés de la décomposition de Owen , si elle sont considérées comme « jouant ensemble » tout en illustrant l'effet égalisateur de lmg : les poids des deux variables sont plus similaires avec lmg qu'avec pmvd.

	lmg 2009	**pmvd 2009**	**lmg 2015**	**pmvd 2015**
Education	31,5 %	56,3 %	45,9 %	64,0 %
Examination	25,6 %	1,0 %	38,7 %	23,5 %
Total	57,1 %	57,3 %	84,6 %	87,5 %

Tableau 4.4.1.3. % d'importance *lmg-Shapley/pmvd.* Données quadratisées (2009) ou non (2015).

Les calculs réalisés après avec RF-CART mettront plus loin en perspective cette comparaison des résultats avec ou sans transformation quadratique.

En effet, la comparaison des valeurs *lmg-Shapley* et RF-CART avec mtry=1 sur données non quadratisées (lmg 2015) montre que RF avec mtry=1 sur ces données non quadratisées donne des résultats plus proches de lmg sur ces données une fois quadratisées (2009).

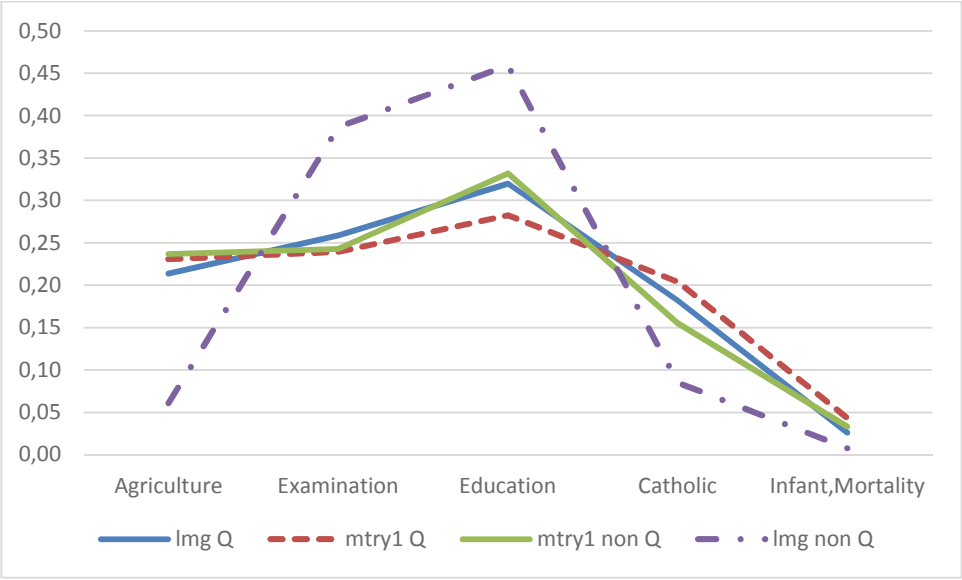


Figure 4.4.1.3 RF-mtry=1 et *lmg-Shapley* sur données *swiss 182* quadratisées (Q) et non quadratisées (non Q).

Ceci suggère dans cet exemple particulier que RF-CART avec mtry=1 est bien à même de prendre en compte les non linéarités et de fournir une mesure plus convaincante que lmg.

Le tableau de synthèse ci-après regroupe les importances normalisées à 100 % calculées avec trois options :

- *swiss* 182 quadratisées et ntree=250
- *swiss* 182 quadratisées et ntree=2000
- *swiss* 182 non quadratisées et ntree=250

Quadratic ntree=250											
	lmg	pmvd	last	first	betasq	pratt	mtry1	mtry2	mtry3	mtry4	mtry5
Agriculture	21,36%	20,54%	44,83%	14,53%	32,99%	26,27%	22,29%	25,27%	25,66%	27,41%	29,27%
Examination	25,85%	1,06%	0,25%	32,71%	0,56%	5,13%	24,24%	21,27%	19,33%	15,82%	16,83%
Education	32,00%	56,34%	24,90%	33,36%	40,87%	44,31%	27,87%	32,08%	34,59%	35,03%	34,60%
Catholic	18,19%	18,57%	20,26%	19,21%	19,00%	22,93%	19,14%	17,49%	16,12%	16,72%	14,52%
Infant,Mortality	2,60%	3,49%	9,76%	0,19%	6,58%	1,36%	6,46%	3,89%	4,31%	5,03%	4,79%
Σ	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Quadratic ntree=2000											
	lmg	pmvd	last	first	betasq	pratt	mtry1	mtry2	mtry3	mtry4	mtry5
Agriculture	21,36%	20,54%	44,83%	14,53%	32,99%	26,27%	23,07%	24,61%	25,60%	27,40%	28,89%
Examination	25,85%	1,06%	0,25%	32,71%	0,56%	5,13%	23,95%	19,90%	19,21%	16,12%	15,19%
Education	32,00%	56,34%	24,90%	33,36%	40,87%	44,31%	28,25%	33,39%	33,84%	35,38%	35,15%
Catholic	18,19%	18,57%	20,26%	19,21%	19,00%	22,93%	20,41%	18,11%	17,25%	16,59%	16,48%
Infant,Mortality	2,60%	3,49%	9,76%	0,19%	6,58%	1,36%	4,32%	3,99%	4,11%	4,52%	4,29%
Σ	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Non Quadratic n=250											
	lmg	pmvd	last	first	betasq	pratt	mtry1	mtry2	mtry3	mtry4	mtry5
Agriculture	6,10%	4,84%	16,22%	5,84%	7,05%	6,85%	23,67%	25,47%	24,76%	28,35%	28,88%
Examination	38,68%	23,51%	9,58%	41,44%	13,53%	25,29%	24,28%	22,43%	27,38%	19,09%	22,71%
Education	45,94%	63,99%	61,87%	42,26%	72,73%	59,22%	33,20%	38,06%	35,98%	43,00%	38,36%
Catholic	8,52%	7,08%	9,80%	10,22%	5,61%	8,09%	15,52%	10,72%	9,13%	7,20%	7,92%
Infant,Mortality	0,76%	0,58%	2,54%	0,25%	1,09%	0,55%	3,32%	3,31%	2,76%	2,35%	2,12%
Σ	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tableau 4.4.1.4. Impact de la quadratisation et de ntree. Données *swiss* 182.

Ces résultats appellent les observations suivantes :

- sur les données « Non Quadratic » les valeurs calculées ici pour *lmg, pmvd, last, first betasq* et *pratt* correspondent exactement aux résultats publiés par Grömping (2015) et sont aussi sur les données « Quadratic » très proches des résultats publiés par Grömping en 2009. Ceci confirme l'exactitude des calculs effectués.

- sur les données « Non Quadratic » Les valeurs obtenues par RF mtry=1 ne sont plus comme dans d'autres exemples aussi proches des lmg.

- avec mtry1, RF-CART donne des résultats relativement proches de lmg 2009 c'est-à-dire de lmg sur variables quadratisées, et ces résultats sont proches à la fois avec l'utilisation des jeux de données quadratisées et non quadratisées.

Ceci renvoie à la capacité des forêts aléatoires à prendre en compte les non linéarités, la propriété connue des arbres s'étend donc à la forêt. Avec *lmg-Shapley*, les allocations entre variables notamment pour « Agriculture » et « Catholic » changent nettement selon que les variables sont quadratisées ou non, tandis que RF-CART avec mtry=1 donne des résultats proches dans les deux cas.

Au 4.3 il a été constaté que lmg-Shapley et RF avec mtry=1 tendaient à être assez proches. L'analyse supplémentaire des données *swiss* 182 confirme la capacité de RF-CART à tenir compte de la non-linéarité possible entre les prédicteurs et la variable à prédire. Cette propriété des arbres se manifeste de façon tangible au niveau de la forêt aléatoire et assez logiquement la proximité la plus forte est réalisée avec la randomisation maximale dans le cas des données *swiss* 182. Ce résultat conduit à renforcer l'appréciation de l'intérêt de l'approche par les forêts aléatoires par rapport aux conclusions du panorama des méthodes présentées par Grömping (2015).

4.4.2 Simulations avec données quadratisées

De façon analogue avec les simulations effectuées au 3.1.12 un script R a été écrit pour générer un jeu composé d'une variable à prédire Y et de 6 prédicteurs $X_1,...,X_6$, en choisissant arbitrairement la matrice de corrélation entre ces sept variables.

Une nouvelle variable à prédire est alors générée Y_{sq} en choisissant notamment des combinaisons linéaires quadratiques en X_1 et X_2.

La variable simulée Y_{sq} est générée de la façon suivante :

$$Y_{sq} = \alpha_1 X_1 + \alpha_2 X_1^2 + \beta_1 X_2 + \beta_2 X_2^2 + \delta X_6 + kY + \gamma N_{norm}$$

Les paramètres α_1 , α_2 , β_1 , β_2 permettent de choisir le degré de quadratisation de la nouvelle variable à prédire en fonction de X_1 et X_2. Le paramètre k permet de choisir la force de la liaison avec la variable Y générée

initialement. Le paramètre γ permet d'ajouter une variable de bruit, V_{norm} étant une variable normale standardisée et le paramètre δ de faire rentrer une variable supplémentaire X_6 dans la construction de Y_{sq} .

Afin d'étudier la manière dont la non linéarité est traitée la simulation comprend les étapes suivantes :

1. Dans un premier temps, Y_{sq} est décomposée avec les méthodes précédentes y compris avec les forêts aléatoires sur l'ensemble des mtry possibles en utilisant les 6 variables $X_1,...X_6$.

2. Dans un deuxième temps les variables X_1 et X_2 sont quadratisées comme effectué au 4.4.1 c'est-à-dire que la variable à prédire est régressée sur les huit variables (les six de départs puis les deux variables X_1 et X_2 élevées au carré) et les variables X_1 et X_2 sont alors remplacées par les combinaisons linéaires de chacune de ces variables et de leur carrés respectifs en utilisant les coefficients de la régression effectuée sur les huit prédicteurs.

3. Y_{sq} est alors décomposé à nouveau selon les méthodes précédentes. Comme attendu la régression OLS de Y_{sq} génère bien par construction des coefficients β exactement égaux à 1 pour les deux nouveaux prédicteurs X_1 et X_2 ainsi quadratisés.

4. Les résultats reportés ci-après pour chaque configuration simulée sont les R^2 avant et après quadratisation, ainsi que la comparaison des résultats de lmg-Shapley et RF-CART avec mtry=1 avant et après quadratisation, afin de vérifier la capacité de RF-CART avec mtry=1 de donner des résultats cohérents selon que les variables sont ou non quadratisées, et de voir comment en revanche évoluent les décompositions de lmg-Shapley.

Pour une première série de simulation en fonction des paramètres la matrice de corrélation des 7 prédicteurs initiaux est choisie comme suit :

Cor	Y	X1	X2	X3	X4	X5	X6
Y	1,00	0,70	0,50	0,30	0,10	0,15	0,13
X1	0,70	1,00	0,20	0,15	0,10	0,05	0,10
X2	0,50	0,20	1,00	0,80	0,10	0,05	0,07
X3	0,30	0,15	0,80	1,00	0,10	0,12	0,10
X4	0,10	0,10	0,10	0,10	1,00	0,40	0,35
X5	0,15	0,05	0,05	0,12	0,40	1,00	0,28
X6	0,13	0,10	0,07	0,10	0,35	0,28	1,00

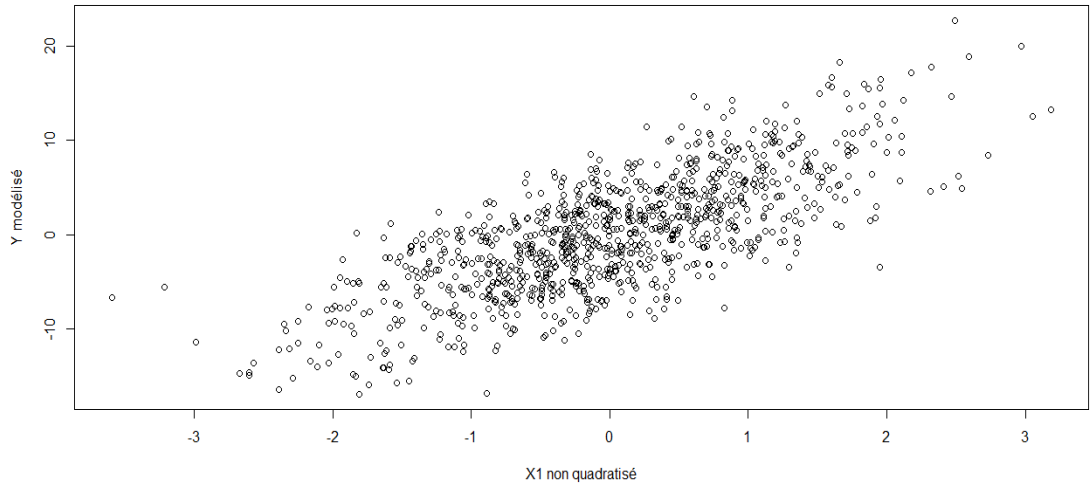
Dans cette configuration X_1, X_2 et X_3 sont classées en ordre décroissant de corrélation avec a variable à prédire.

Simulation A

Pour cette simulation les valeurs des paramètres sont les suivantes :

$\alpha_1 = 1 \quad \alpha_2 = 0 \quad \beta_1 = 1 \quad \beta_2 = 0 \quad \gamma = 1 \quad \delta = 0 \quad k = 5,$

La relation entre Y_{sq} et X_1 avant quadratisation est illustrée ci-dessous.



Les R^2 avant quadratisation et après quadratisation sont presque identiques : 0,779 et les quatre résultats (lmg et RF-CART mtry=1 avant et après quadratisation) sont aussi presque identiques :

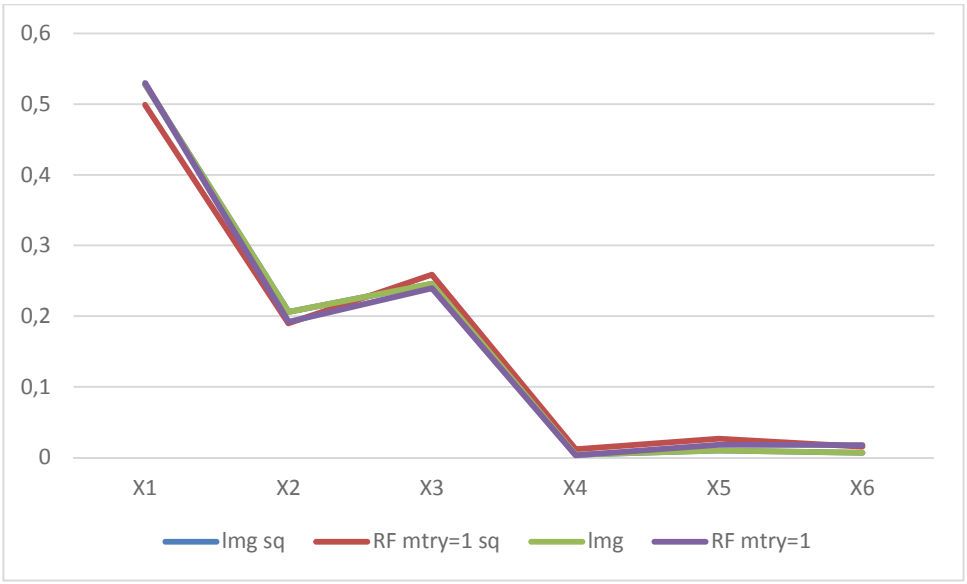


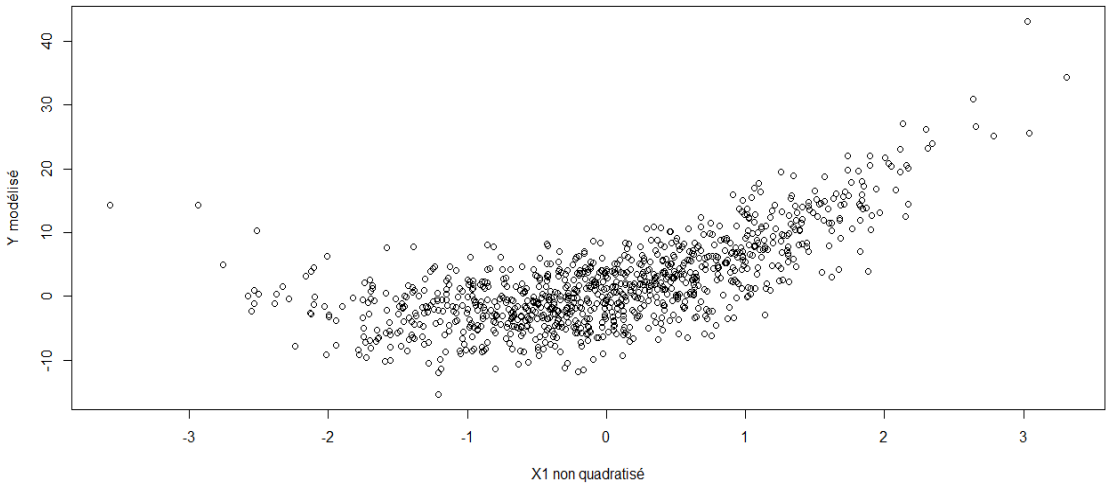
Figure 4.4.2.1. Données simulées. Impact de la quadratisation de X_1 et X_2.

Simulation B

Pour cette simulation les valeurs des paramètres sont les suivantes :

$$\alpha_1 = 1 \quad \alpha_2 = 2 \quad \beta_1 = 1 \quad \beta_2 = 0 \quad \gamma = 1 \quad \delta = 0 \quad k = 5,$$

La relation entre Y_{sq} et X_1 avant quadratisation est illustrée ci-dessous montrant cette fois une légère quadratisation.



Le R^2 avant quadratisation est 0,6435 et est 0,8178 après quadratisation. Les quatre résultats sont présentés ci-dessous :

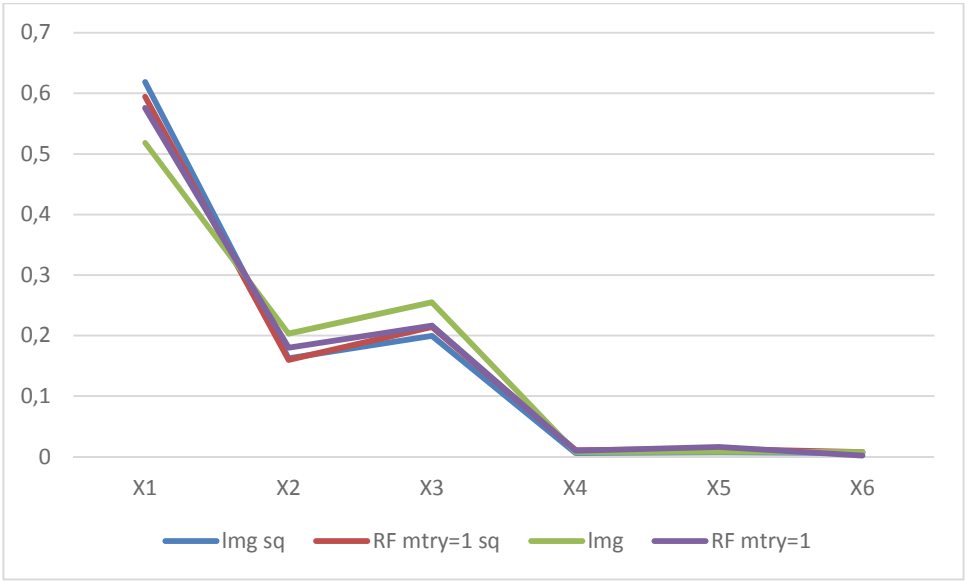


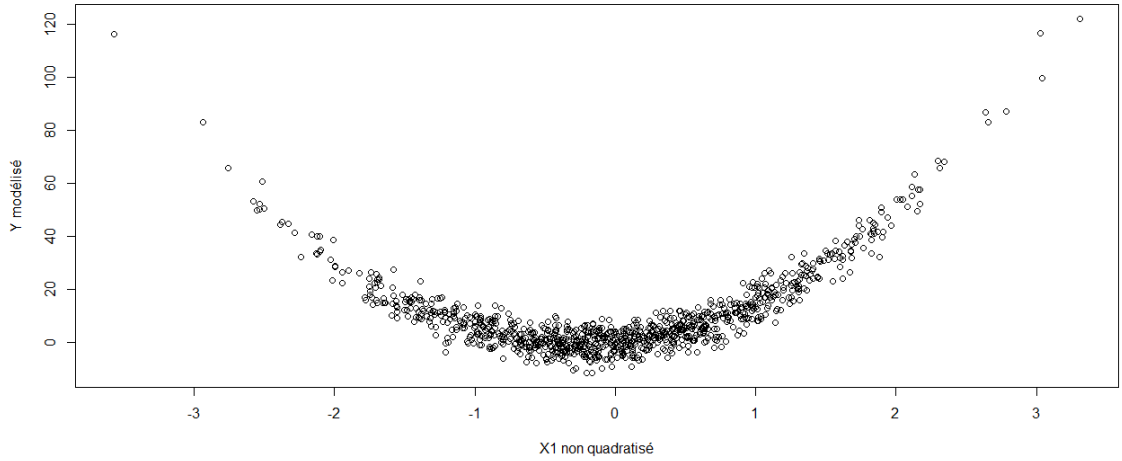
Figure 4.4.2.2. Données simulées. Impact de la quadratisation de X_1 et X_2.

La légère quadratisation introduite se manifeste dans une différence des résultats pour lmg-Shapley selon que les variables sont ou non quadratisées, tandis que les deux résultats avec RF-CART mtry=1 restent très semblables.

Simulation C

Pour cette simulation les valeurs des paramètres sont les suivantes :

$\alpha_1 = 1 \quad \alpha_2 = 10 \quad \beta_1 = 1 \quad \beta_2 = 0 \quad \gamma = 1 \quad \delta = 0 \quad k = 5$. La relation entre Y_{sq} et X_1 avant quadratisation cette fois nettement plus quadratique.



Le R^2 avant quadratisation est 0,132 et est 0,963 après quadratisation. Les quatre résultats sont ci-dessous :

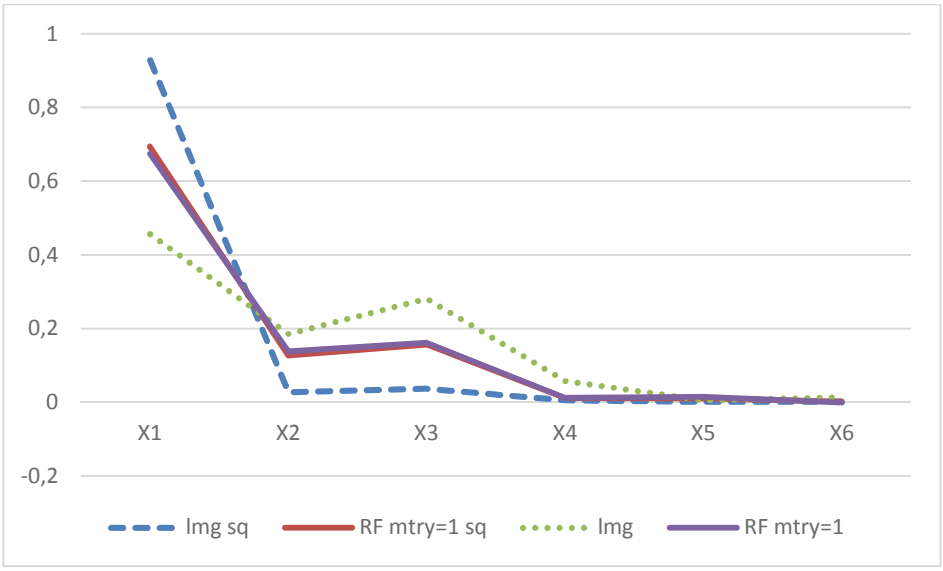


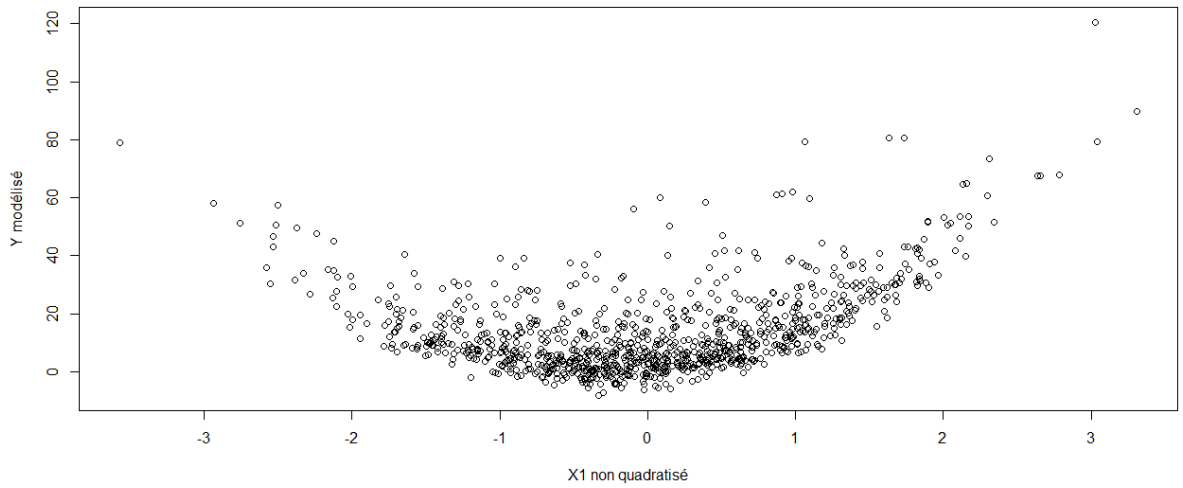
Figure 4.4.2.2. Données simulées. Impact de la quadratisation de X_1 et X_2.

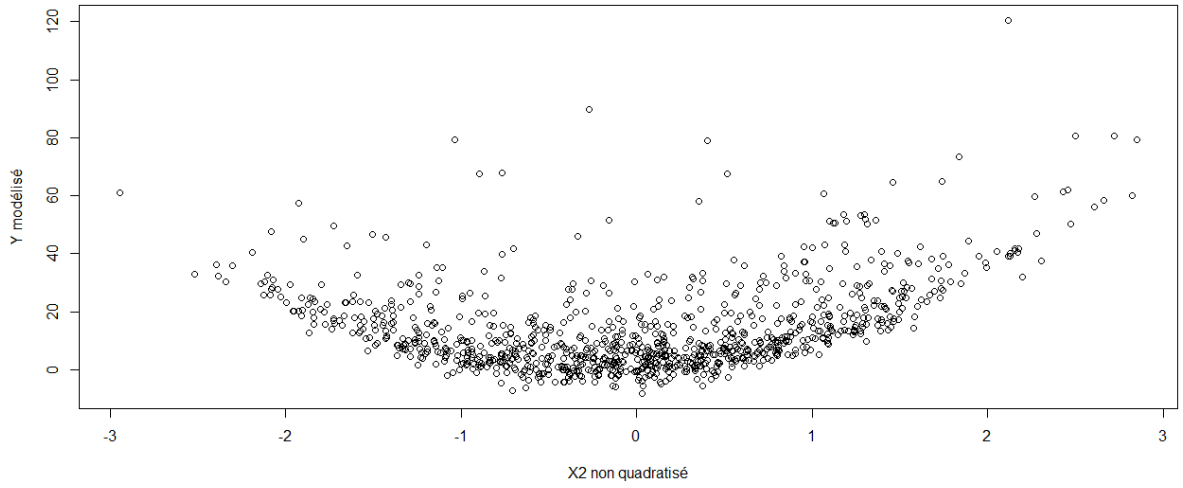
La quadratisation plus forte introduite pour la variable X_1 se manifeste dans une différence des résultats pour lmg-Shapley accrue selon que les variables sont ou non quadratisées, tandis que les deux résultats avec RF-CART mtry=1 restent une nouvelle fois très semblables. Relevons que dans cette configuration d'effet quadratique très prononcé pour X_1 l'importance en % est presque toute allouée à ce prédicteur avec lmg-Shapley sur les variables non quadratisées.

Simulation D

Pour cette simulation les valeurs des paramètres sont les suivantes afin de combiner des quadratisations à la fois sur X_1 et X_2.

$\alpha_1 = 1$ $\alpha_2 = 7$ $\beta_1 = 1$ $\beta_2 = 7$ $\gamma = 1$ $\delta = 0$ $k = 5$. La relation entre Y_{sq} et X_1 avant quadratisation cette fois intermédiaire entre les simulations B et C présentées précédemment.





Le R^2 avant quadratisation est 0,146 et est 0,961 après quadratisation. Les quatre résultats sont présentés ci-dessous :

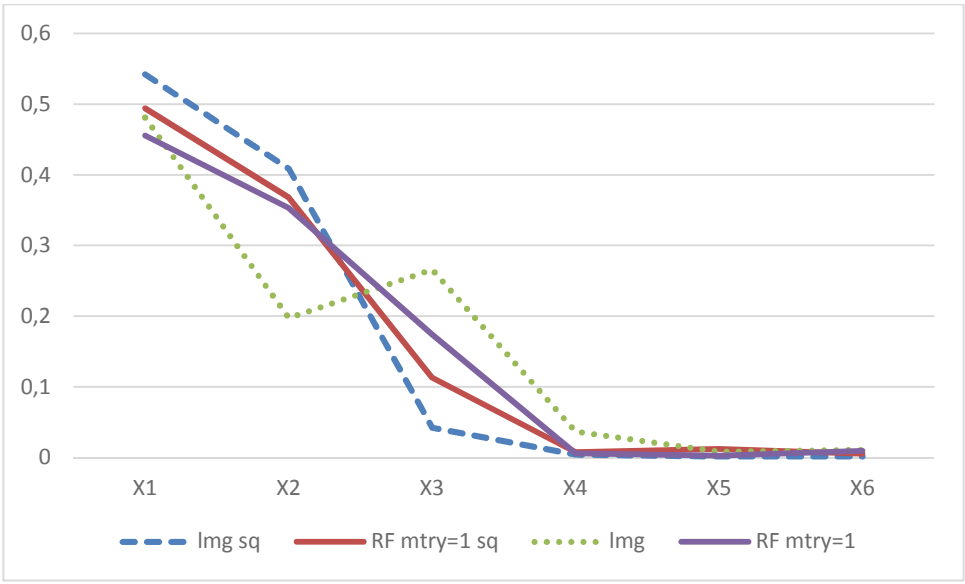


Figure 4.4.2.3. Données simulées. Impact de la quadratisation de X_1 et X_2

La quadratisation forte introduite pour les variables X_1 et X_2 se manifeste dans une différence des résultats pour lmg-Shapley selon que les variables sont ou non quadratisées, tandis que les deux résultats avec RF-CART mtry=1 restent une nouvelle fois très semblables.

Les corrélations avec Y_{sq} sont les suivantes (en notant X_{1q} et X_{2q} les variables après la quadratisation) :

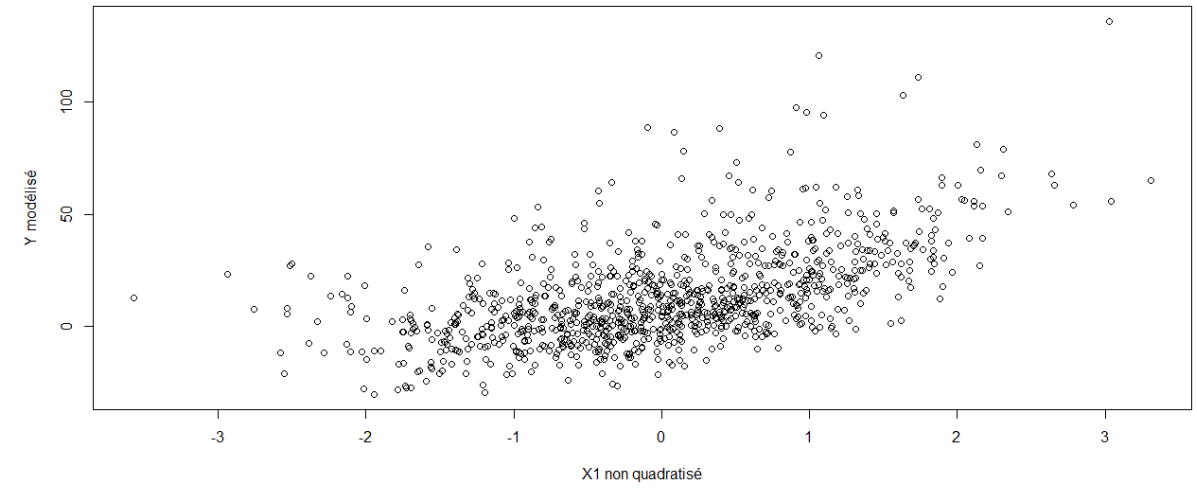
$$cor(Y_{sq}, X_1) = 0,316 \qquad cor(Y_{sq}, X_{1q}) = 0,746$$
$$cor(Y_{sq}, X_2) = 0,255 \qquad cor(Y_{sq}, X_{2q}) = 0,653$$
$$cor(Y_{sq}, X_3) = 0,293$$

Par rapport à RF-CART mtry=1, *lmg-Shapley* va allouer une importance plus faible à X_2 avant quadratisation, ce qui peut s'interpréter comme la conséquence de ce que *lmg-Shapley* ne tient pas compte de la relation quadratique entre Y_{sq} et X_2. D'un autre côté après quadratisation *lmg-Shapley* renforce bien les allocations pour X_1 et X_2 mais cette fois va allouer une importance relative beaucoup plus faible à X_3.

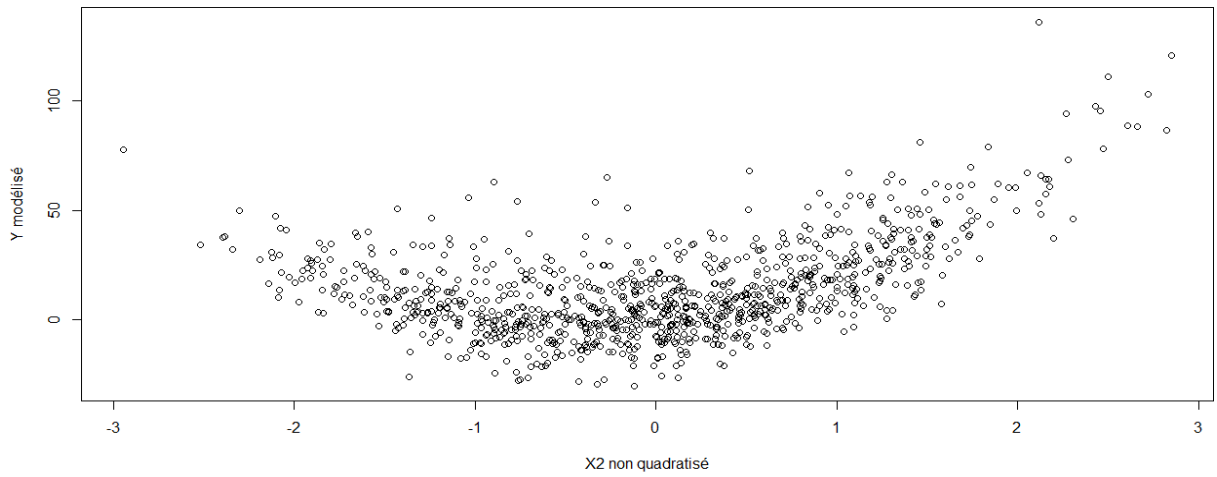
Simulation E

Pour cette simulation la différence avec la simulation D ci-dessus va être d'augmenter le facteur k c'est-à-dire de rapprocher Y_{sq} du prédicteur simulé initialement Y .

$\alpha_1 = 1 \quad \alpha_2 = 3 \quad \beta_1 = 1 \quad \beta_2 = 10 \quad \gamma = 1 \quad \delta = 0 \quad k = 15$. La relation entre Y_{sq} et X_1 sera donc légèrement quadratique mais moins que celle entre Y_{sq} et X_2 comme illustré ci-dessous :



.



Le R^2 avant quadratisation est 0,421 et est 0,827 après quadratisation. Les quatre résultats sont présentés ci-dessous :

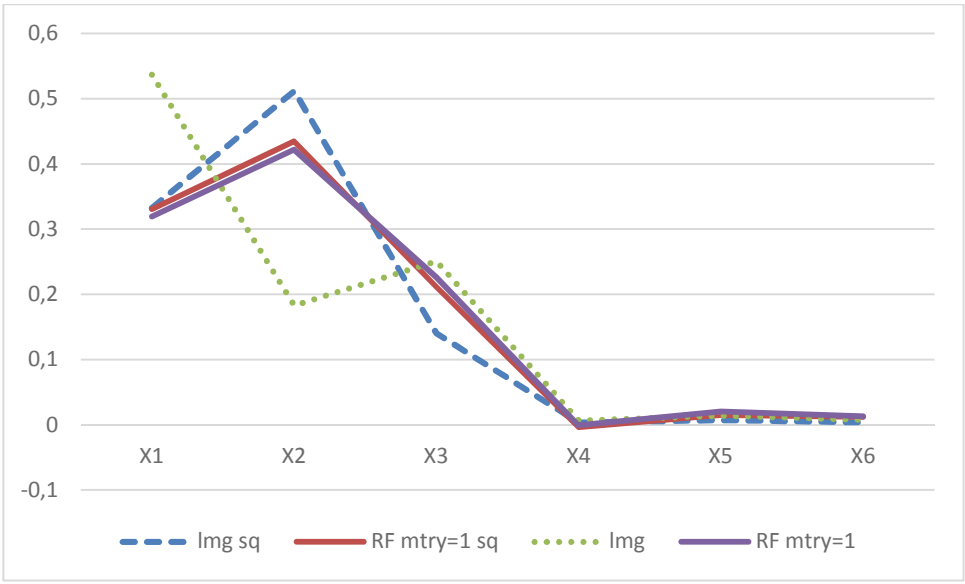


Figure 4.4.2.4. Simulation « E ».

La quadratisation forte introduite pour les variables X_1 et X_2 se manifeste dans une différence des résultats pour lmg-Shapley selon que les variables sont ou non quadratisées, tandis que les deux résultats avec RF-CART mtry=1 restent une nouvelle fois très semblables.

Les corrélations avec Y_{sq} sont les suivantes (en notant X_{1q} et X_{2q} les variables après la quadratisation) :

$$cor(Y_{sq}, X_1) = 0,556 \qquad cor(Y_{sq}, X_{1q}) = 0,569$$
$$cor(Y_{sq}, X_2) = 0,417 \qquad cor(Y_{sq}, X_{2q}) = 0,705$$
$$cor(Y_{sq}, X_3) = 0,456$$

Par rapport à RF-CART mtry=1 lmg-Shapley va allouer ici encore une importance plus faible à X_2 avant quadratisation, ce qui peut s'interpréter comme la conséquence de ce que *lmg-Shapley* ne tient pas compte de la relation quadratique entre Y_{sq} et X_2. X_2 ayant une relation très quadratique avec Y_{sq} « gagne » beaucoup d'importance après quadratisation comme le montre également la corrélation bivariée entre Y_{sq} et X_{2q} nettement plus élevé qu'entre Y_{sq} et X_2.

lmg-Shapley accorde plus d'importance avant quadratisation à X_1 tandis que RF-CART avec mtry=1 pondère de façon cohérente avant et après quadratisation entre X_1 et X_2

Ces différentes simulations confirment toutes la capacité de la méthode des forêts aléatoires à prendre en compte la non linéarité des relations entre les prédicteurs et la variable à prédire dans l'estimation de l'importance, et à donner des résultats cohérents avant et après quadratisation. De ce point de RF-CART avec mtry=1 est plus stable et est en

outre apte à prendre en compte les non linéarités quand il s'agit d'attribuer une importance relative que ne le fait lmg-Shapley.

Simulation F

Grömping (2009) étudie sur des simulations avec 4 prédicteurs l'impact du changement de mtry sur les allocations en utilisant RF-CART et indique qu'elles sont très modérées et moindres qu'avec le package RF-CI.
Dans des simulations avec plus de prédicteurs permettant une évolution plus large de mtry (7 au lieu de 4) et en fixant une corrélation forte entre certains prédicteurs, il apparaît en fait une variabilité plus grande des allocations en accroissant mtry. Ceci est illustré dans la simulation ci-dessous :

Matrice de corrélations :

	Ysq	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X1eps**
Ysq	1,00	0,56	0,51	0,43	0,03	0,11	0,09	0,50
X1	0,56	1,00	0,80	0,35	0,10	0,05	0,10	0,89
X2	0,51	0,80	1,00	0,80	0,10	0,05	0,07	0,71
X3	0,43	0,35	0,80	1,00	0,10	0,12	0,10	0,31
X4	0,03	0,10	0,10	0,10	1,00	0,40	0,35	0,09
X5	0,11	0,05	0,05	0,12	0,40	1,00	0,28	0,04
X6	0,09	0,10	0,07	0,10	0,35	0,28	1,00	0,08
X1eps	0,50	0,89	0,71	0,31	0,09	0,04	0,08	1,00

Le R^2 est 0,4851. X_{1eps} est une variable bruitée construite à partir de X_1 en ajoutant une variable aléatoire normale standardisée de variance 0,3.

L'évolution des allocations en fonction de mtry est illustrée ci-dessous :

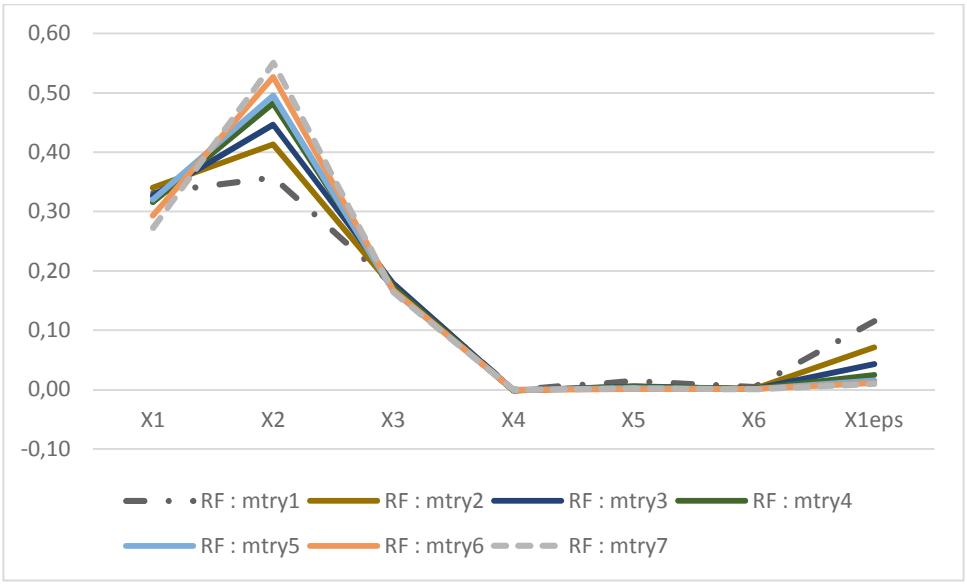


Figure 4.4.2.5. Simulation « F ».

Cette simulation particulière monte une plus grande influence de mtry sur les allocations par RF-CART.

4.5 Sélection de Variables avec VSURF

Depuis les travaux de Grömping, des développements ont été réalisés sur la sélection de variables en utilisant les Forêts aléatoires (Genuer, Poggi, Tuleau-Malot (2008, 2010)) et un package R a été déposé par ces auteurs (VSURF : Variable Selection Using Random Forest).

VSURF comprend plusieurs étapes. Tout d'abord les variables sont classées par ordre décroissant d'importance (moyenne sur 50 calculs). Ensuite les calculs se déroulent ainsi :

- élimination de variables superflues avec l'utilisation d'un seuil minimum. Ce seuil est déterminé en construisant d'abord une courbe où les variables en abscisse sont ordonnées comme indiqué ci-dessus et où en ordonnées sont figurées les écart-types des importances MSE. Le seuil d'élimination des variables correspond à la valeur minimum de prédiction d'un CART ajusté à cette courbe. Seules les variables classées avant la variable correspondant à ce seuil sont gardées pour les étapes de calculs suivantes.

- sélection parmi les variables ainsi retenues d'un ensemble de variables explicatives en vue d'une interprétation. Les erreurs OOB sont calculées en prenant successivement les variables non éliminées par ordre d'importance MSE décroissantes, seules les variables situées avant celle correspondant au minimum d'erreur OOB sont sélectionnées comme les variables pour l'interprétation.

- sélection d'un sous-ensemble restreint de variables en vue d'une prédiction. Les variables retenues à l'étape de sélection pour interprétations sont introduites jusqu'à ce que le gain en OOB erreur soit inférieur à un seuil ainsi défini comme suit :

Soit $p_{\text{int}erp}$ le nombre de variables sélectionnées pour l'interprétation et $p_{e\lim}$ le nombre de variables sélectionnées à la phase d'élimination. En introduisant une à une par ordre décroissant d'importance MSE les variables non sélectionnées pour l'interprétation le seuil est calculé de la façon suivante :

$$\frac{1}{p_{e\lim} - p_{\text{int}erp}} \sum_{j=p_{\text{int}erp}}^{j=p_{e\lim}} \left| OOB(j+1) - OOB(j) \right|$$

Le package VSURF a été appliqué ci-après sur différents jeux de données utilisés précédemment en vue de déterminer et comparer les effets des sélections et aussi de tester différents paramétrages de VSURF, c'est-à-dire le choix de mtry pour la phase d'élimination, puis le choix de mtry pour la sélection en vue de l'interprétation.

Le jeu *swiss* 182 a également été analysé avec VSURF à la fois avec les données quadratisées et non quadratisées. Pour les données non quadratisées, VSURF identifie 3 variables pour l'interprétation et la prédiction :

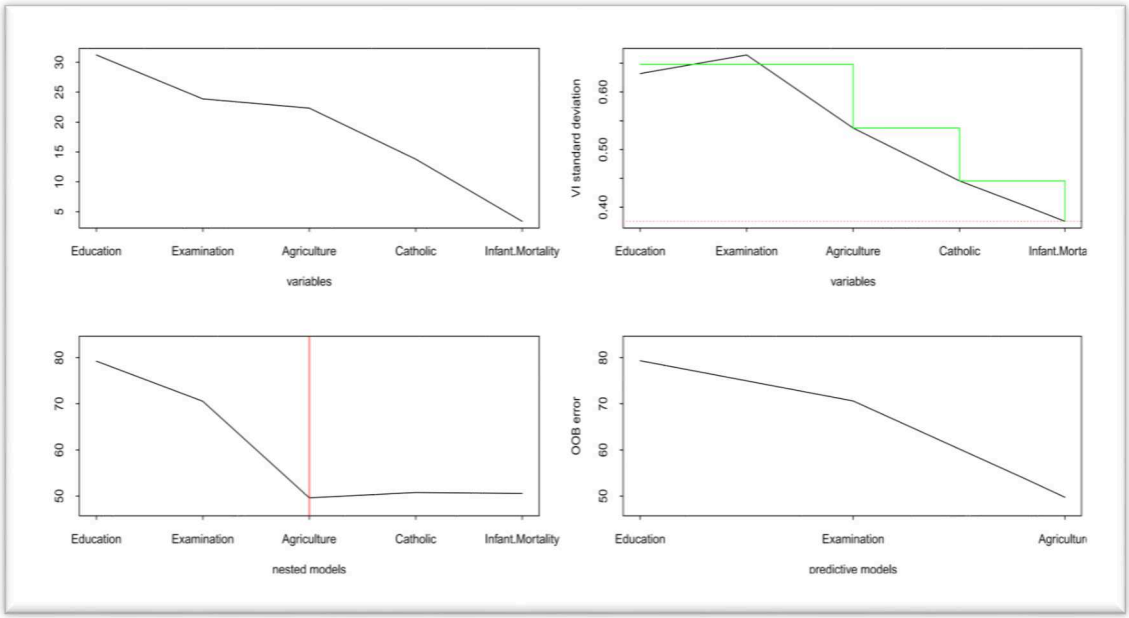


Figure 4.5.1 Sélection avec VSURF. Données *swiss*182 non quadratisées.

Pour les données quadratisées, VSURF ajoute cette fois la variable quadratisée « Catholic » tant au niveau de l'interprétation que de la prévision.

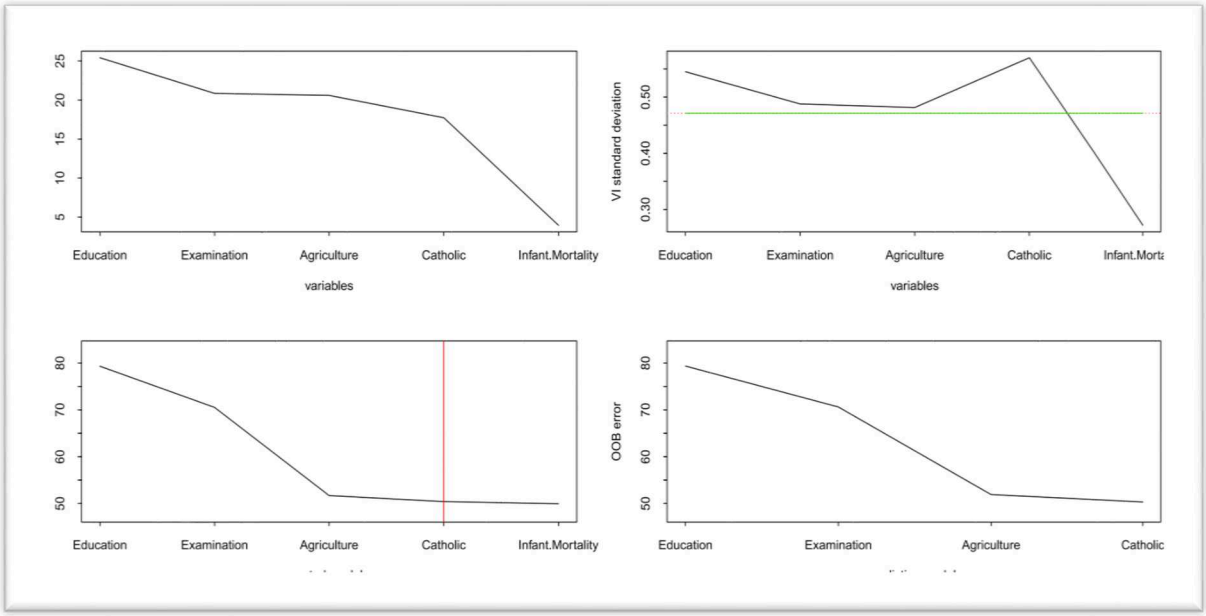


Figure 4.5.2 Sélection avec VSURF. Données *swiss*182 quadratisées.

Voici la comparaison avec les sorties numériques obtenues avec VSURF :

	No	VI Q	VI Non Q	OOB Modèles Q	OOB Modèles Non Q
Education	3	0,29	0,33	**79,30**	**79,26**
Examination	2	0,24	0,25	**70,50**	**70,56**
Agriculture	1	0,23	0,24	**51,70**	**49,62**
Catholic	4	0,20	0,15	**50,40**	50,77
Mortality	5	0,04	0,04	50,00	50,55
Σ		1	1		

Tableau 4.5.1. Sélection des variables avec VSURF. Données *swiss* 182 quadratisées (Q) et non quadratisées (Non Q).

VSURF a également été appliqué sur le jeu *UK Data* en comparant les résultats selon la valeur de mtry, avec mtry=1, mtry=défaut (ici 4) et mtry=14. En ce qui concerne les seuils pour l'élimination, il n'y a aucune élimination ni pour mtry=1, défaut ou mtry=14

Il est possible de paramétrer la phase de sélection pour l'interprétation en utilisant une valeur spécifiée de mtry dans la construction des arbres utilisés pour cette sélection, par exemple spécifier mtry=1 pour l'interprétation :

```
X.vsurf.interp <- VSURF_interp(x=x,y=y,vars=X.vsurf$varselect.thres,mtry=1)
```

ou inversement mtry=14 :

```
X.vsurf.interp <- VSURF_interp(x=x,y=y,vars=X.vsurf$varselect.thres,mtry=14)
```

Nous avons ici utilisé 5 configurations (de gauche à droite dans le tableau ci-après)

- Sélection : mtry=1, Interprétation : Défaut
- Sélection : mtry=1, Interprétation : mtry=1
- Sélection : mtry=Défaut, Interprétation : mtry=Défaut
- Sélection : mtry=14, Interprétation : mtry= Défaut
- Sélection : mtry=14, Interprétation : mtry=14

Les variables sont ordonnées en fonction du résultat de la phase de sélection initiale (qui ne dépend que de mtry) qui est indiqué sur la première ligne du tableau, tandis que sur la deuxième ligne figure le mtry retenu pour l'étape de choix des variables pour l'interprétation. Les variables retenues pour l'interprétation sont figurées en caractère gras dans chaque colonne.

Sélection	**mtry=1**	**mtry=1**		**mtry=D**	**mtry=D**			**mtry=14**	**mtry=14**
Interprétation	**mtry= D**	**mtry=1**			**mtry=D**			**mtry=D**	**mtry =14**
Variables				**Variables**			**Variables**		
14	**1,703**	**1,702**		14	**1,702**		11	**1,818**	**1,818**
11	**1,609**	**1,608**		11	**1,609**		14	**1,609**	**1,612**
13	**1,602**	**1,602**		2	**1,508**		2	**1,508**	**1,536**
2	**1,512**	**1,512**		13	**1,513**		13	**1,512**	1,566
7	**1,504**	**1,504**		5	**1,499**		5	**1,498**	1,577
5	**1,517**	**1,494**		7	1,517		1	1,512	1,619
4	**1,502**	**1,482**		1	1,518		12	1,507	1,609
3	1,512	**1,489**		3	1,528		7	1,505	1,612
6	1,546	**1,489**		12	1,549		3	1,551	1,629
9	1,535	**1,480**		4	1,534		9	1,513	1,574
12	1,506	**1,472**		9	1,505		4	1,506	1,562
8	1,514	1,470		6	1,521		10	1,504	1,544
10	1,500	1,470		8	1,515		6	1,504	1,546
1	1,500	1,472		10	1,500		8	1,499	1,541

Tableau 4.5.2. VSURF. UK Data. Impact des choix de mtry pour sélection-interprétation.

L'utilisation de mtry=1 aux deux étapes, c'est-à-dire sélection et interprétation conduit à une sélection de 11 variables au lieu de 7 si la phase interprétation est effectuée avec la valeur par défaut.

L'utilisation de mtry =14 aux deux étapes, c'est-à-dire sélection et interprétation conduit à une sélection de 3 variables au lieu de 5 si la phase interprétation est effectuée avec la valeur par défaut. Ces sélections résultent directement des MSE calculées et de la dispersion de la valeur minimale.

Le choix de la valeur par Défaut de mtry est donc cohérent avec cet objectif car il assure pour les valeurs suffisamment grandes du nombre de variables du sous-modèle une erreur OOB minimisée, permettant de ne pas « trop « sélectionner (cas avec mtry=1 aux deux étapes) ni de ne pas sélectionner « trop peu » (cas avec mtry=14 aux deux étapes.

Cette même analyse a été conduite avec le jeu *credit*. (499 observations, 9 prédicteurs).

Les résultats sont présentés ci-après :

		mtry=1			mtry=D			mtry=9	
Variables	**Imp mtry=1**	**err OOB défaut**	**err OOB mtry=1**	**Imp mtry=D**	**err OOB défaut**	**Imp mtry=9**	**err OOB défaut**	**err OOB mtry=9**
7	0,1523	**0,9064**	**0,9061**	0,2291	**0,9917**	0,2602	**0,9914**	**0,9915**
1	0,1541	**0,9652**	**0,9654**	0,2233	**0,9067**	0,2477	**0,9060**	**0,9198**
8	0,1301	0,9095	**0,9108**	0,2004	**0,9104**	0,2356	**0,9103**	0,9692
6	0,1231	0,9153	**0,9151**	0,1720	**0,9151**	0,2052	**0,9146**	1,0260
5	0,1003	0,9363	**0,8969**	0,1610	**0,8975**	0,1904	**0,8983**	1,0013
9	0,1207	0,9106	**0,9109**	0,1566	0,9371	0,1718	0,9377	0,9953
2	0,0972	0,9375	**0,8962**	0,1412	0,9387	0,1676	0,9390	0,9946
3	0,0742	0,9416	0,8987	0,0798	0,9413	0,0897	0,9409	0,9948
4	0,0829	0,9229	**0,8929**	0,0761	0,9415	0,0773	0,9426	0,9717

Tableau 4.5.3. Impact du choix de mtry pour l'étape de sélection VSURF. Données *credit.*

Les outils de R permettent de visualiser le phénomène de changement des sélections pour l'interprétation :

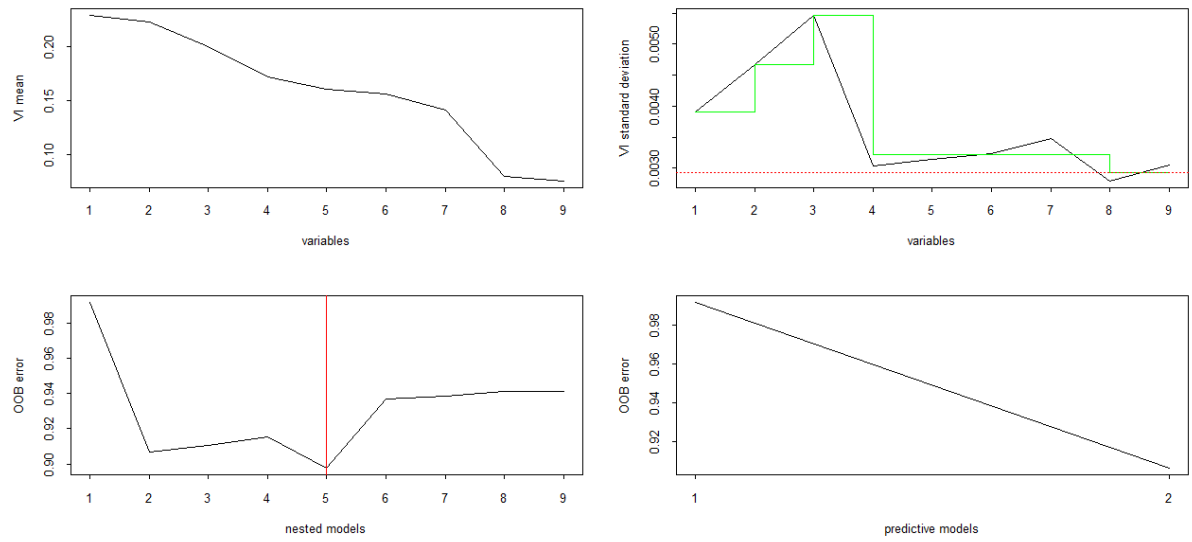


Figure 4.5.3. Sélection avec VSURF. mtry= Défaut aux deux étapes. Données *credit.*

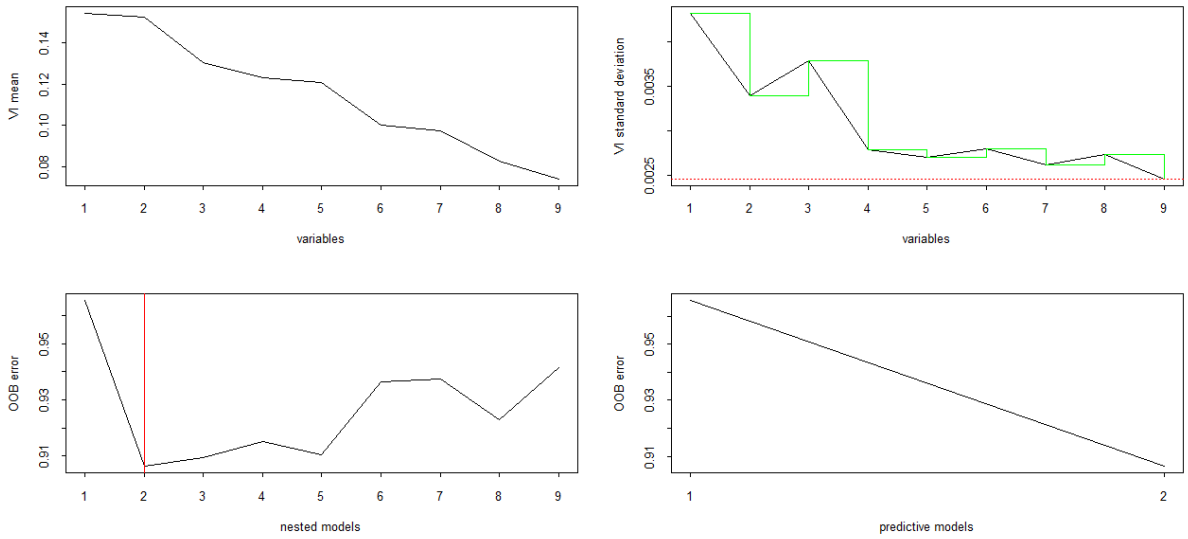


Figure 4.5.4. Sélection avec VSURF.mtry=1 pour sélection. Défaut pour interprétation. Données *credit.*

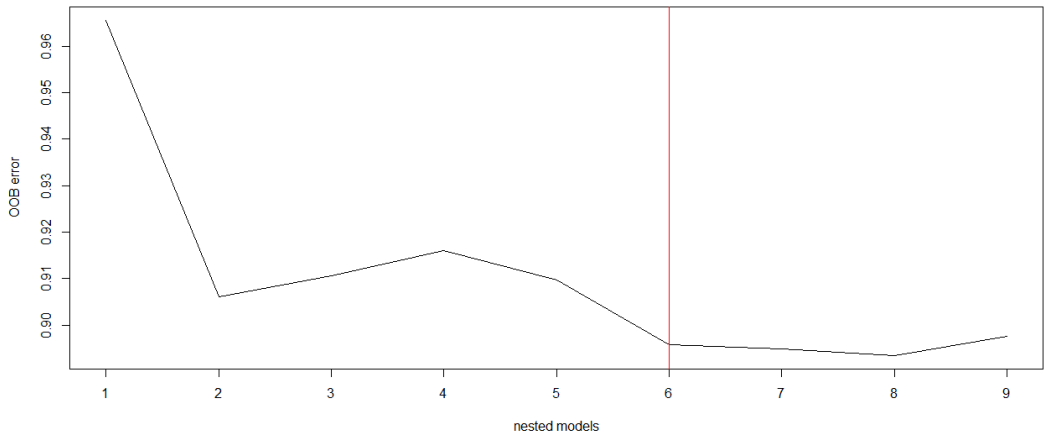


Figure 4.5.5. Sélection avec VSURF. mtry=1 pour sélection. mtry= 1 pour interprétation. Données *credit.*

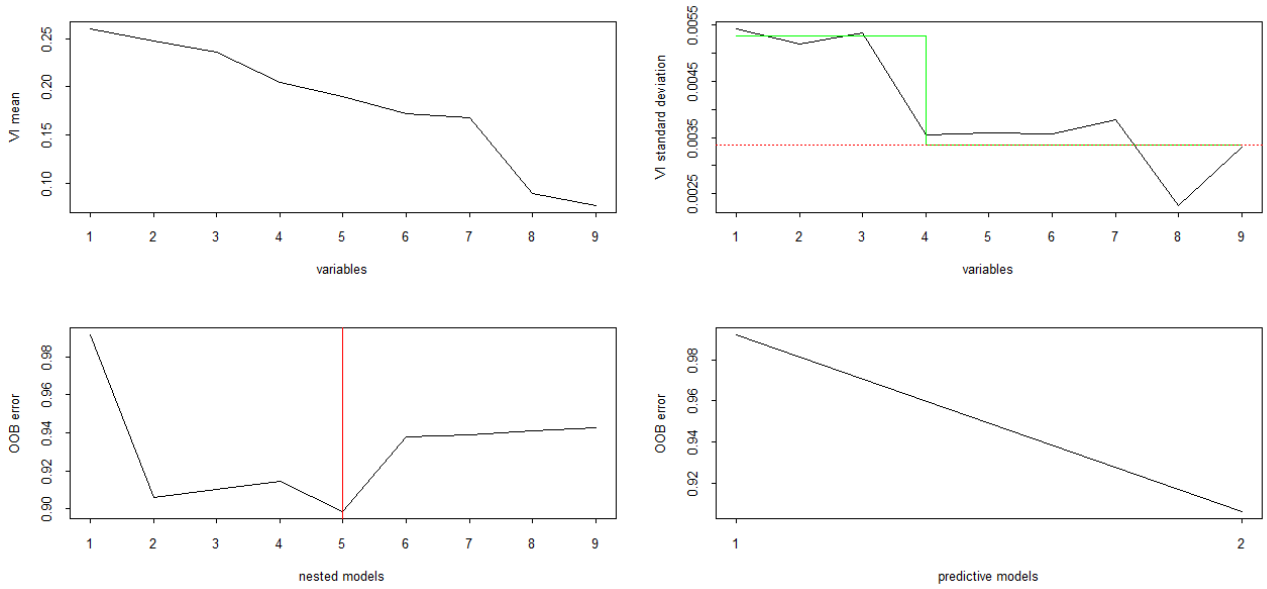


Figure 4.5.6. Sélection avec VSURF. mtry=9 pour sélection. Défaut pour interprétation. Données *credit.*

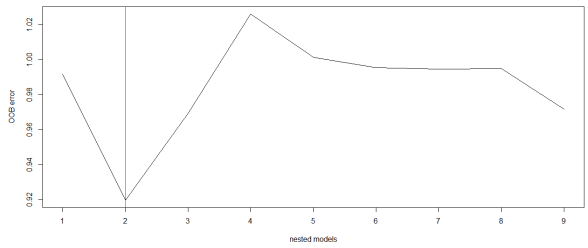


Figure 4.5.7 Sélection avec VSURF. mtry=9 pour sélection et pour interprétation. Données *credit.*

Comme dans le cas des données UK Data, le jeu de données *credit* montre que le recours à un mtry =1 pour le calcul des OOB des sous-modèle accroît la sélection de variables pour l'interprétation tandis qu'au contraire le recours à mtry=p produit l'effet inverse. L'option de combiner deux niveaux de mtry différents pour l'étape de sélection initiale puis pour l'étape de sélection pour l'interprétation ne semble donc pas apporter de bénéfice et en raison des variations sensibles qu'elle introduit dans le processus de sélection pour l'interprétation.

Comme les résultats de RF-CART restent proches entre l'utilisation de mtry=1 ou de la valeur par défaut $p/3$, il apparaît donc plus cohérent pour la sélection de variable d'utiliser les valeurs par défaut aux deux étapes.

En l'absence de sélection de variable si une simple décomposition de la variance est recherchée, RF-CART avec mtry=1 est une solution préférable à lmg-Shapley compte-tenu de sa capacité à prendre en compte les non-linéarités. Cette option permet une sélection la plus aléatoire à chaque étape de construction des arbres, sans pour autant courir de risque important de construire des arbres avec trop de variables de bruit car dans le cas des études de marché le nombre de variables de bruit est en principe nettement inférieur à 50% ce qui garantit que même avec un mtry=1 la probabilité de sélectionner un prédicteur influent est supérieure à 0,5 critère mis en avant par Tufféry (2015).

4.6 Remarques sur les temps de calcul.

Sur la question des temps de calcul Grömping a pris soin de développer le package relaimpo en utilisant la matrice de covariance pour calculer ensuite les valeurs d'importance, ce qui fait que le programme n'est pas sensible au nombre des observations, car une fois les covariances calculées sur l'ensemble des observations, le reste des calculs ne se fait plus sur l'ensemble de l'échantillon. Mais le temps de calcul croit cependant avec le nombre de prédicteurs : doublement pour *lmg-Shapley* avec l'ajout d'un prédicteur et plus que doublement pour pmvd. Grömping (2006) présente l'exemple des temps de calculs suivants (en secondes) :

	100 obs.		1000 obs.	
p	lmg	pmvd	lmg	pmvd
3	0,02	0,02	0,02	0,02
4	0,03	0,03	0,03	0,03
5	0,06	0,05	0,06	0,05
6	0,10	0,09	0,10	0,09
7	0,18	0,18	0,19	0,18
8	0,33	0,37	0,33	0,37
9	0,64	0,78	0,64	0,78
10	1,25	1,74	1,23	1,72
11	2,46	4,22	2,44	4,22
12	4,93	11,64	4,92	11,64

Tableau 4.6.1. Evolution des temps de calculs. Comparaison *lmg-pmvd*. Grömping (2006).

Dans cette situation le temps de calcul pour lmg avec 20 prédicteurs serait de l'ordre de 25 minutes…et dans le cas de PMVD les temps de calculs prohibitifs comme constaté sur des jeux comportant un nombre substantiel de prédicteurs (21 prédicteurs, temps dépassant plusieurs heures…).

Sur le jeu de données *Tcom* (250 observations, 23 prédicteurs) les temps de calculs des importances *first, last* ou *genizi* sont très courts, de l'ordre de quelques centièmes de seconde indépendamment du nombre de prédicteurs. Ils sont aussi très rapides pour RF-CART (mtry=1) et restent de l'ordre de quelques dixièmes de seconde à environ une seconde même avec 20 prédicteurs.

Pour l'importance lmg-Shapley le doublement du temps de calcul avec chaque nouveau prédicteur est sensible et constaté à partir de 15 prédicteurs comme illustré ci-dessous :

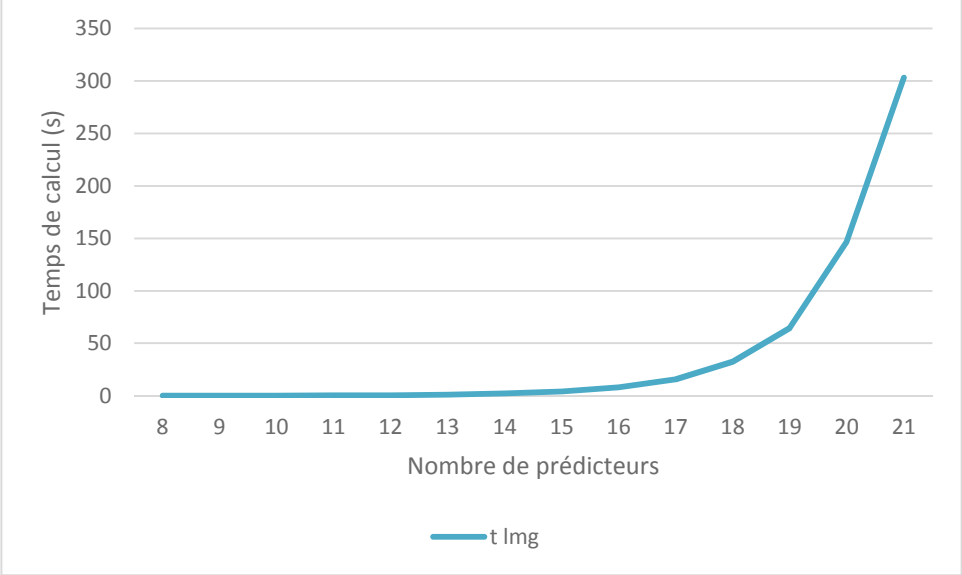


Figure 4.6.1 Temps de calcul *lmg-Shapley*. Données *Tcom*.

En ce qui concerne PMVD, les temps de calculs sont plusieurs centaines de fois plus élevés avec ces mêmes données :

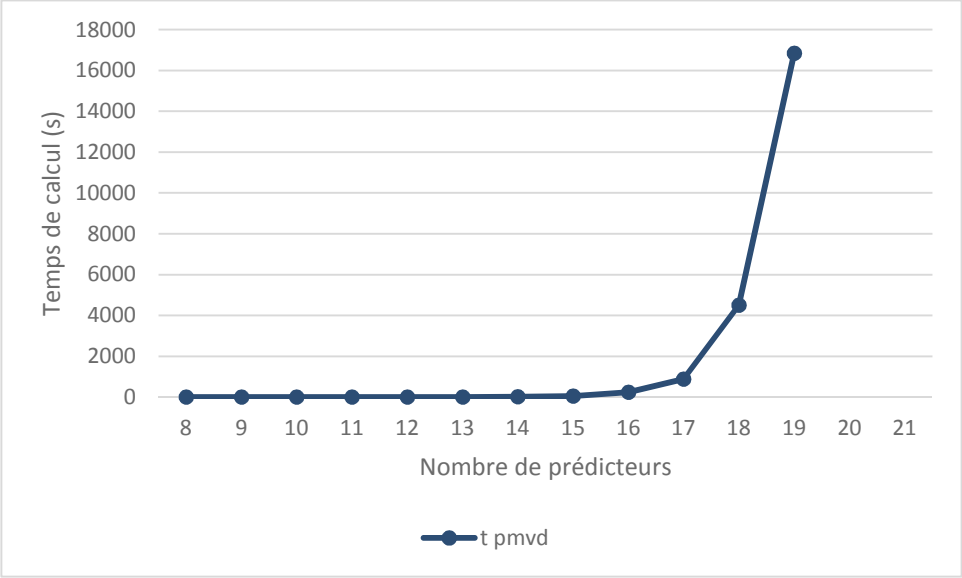


Figure 4.6.2. Temps de calcul *pmvd*. Données *Tcom.*

A titre de commentaire lorsque nous lançons le package `relaimpo` sur un jeu de données avec des prédicteurs presque parfaitement corrélés `relaimpo` retourne un message d'erreur en raison des inversions de matrices alors que weifila ne requiert que les calculs de R2 des modèles et peut donc être programmé pour accepter ces configurations. A contrario `relaimpo` accepte néanmoins sans problème des très fortes colinéarités (ex 0,93 entre des variables de *swiss 182* en ajoutant des variables quadratisées).

Les temps de calculs prohibitifs de *pmvd* rendent son utilisation non réaliste dans les programmes d'études de marchés ou de nombreuses analyses des leviers sont exécutées, comme certains programmes de mesure de la satisfaction à travers de nombreuses lignes de services et territoires.

Cet inconvénient quoique moins fort pour lmg que pour pmvd conduit néanmoins à ne pas privilégier *lmg* par rapport à *genizi*. *weifila* est comme *genizi* très rapide.

Pour un temps de calcul très modéré, RF-CART avec mtry=1 permet en revanche une bien meilleure prise en compte de la non linéarité et des interactions.

Enfin si l'objectif de sélection de variables est recherché, VSURF permet avec des temps de calculs très efficaces (inférieurs à 300 s) de procéder à une sélection, ce que n'offrent pas les autres méthodes.

4.7 Conclusions sur l'apport des forêts aléatoires

Avec les analyses effectuées sur des jeux de données tests ou typiques des applications en études de marchés et aussi les résultats des simulations, les points suivants peuvent être confirmés :

- Il est possible de rencontrer des cas où la valeur de permutation MSE est légèrement négative ce qui va au-delà de la remarque de Grömping (2009) qui mentionne possibilité de cette valeur négative sur un arbre particulier. (Cf. aussi Genuer et al. (2008)).

- Les valeurs de mtry permettant le choix le plus aléatoire entre les prédicteurs à chaque étape de la construction des arbres (mtry=1 ou 2) donnent des résultats proches de lmg-Shapley. Ceci est cohérent avec les caractéristiques de la méthode lmg-Shapley qui tend à la réallocation d'importance entre variables.

- Les valeurs de mtry permettant le choix le moins aléatoire entre les prédicteurs à chaque étape de la construction des arbres (mtry=p ou p-1) donnent des résultats parfois comparables à pmvd mais ces résultats paraissent moins conclusifs que pour la proximité avec lmg-Shapley.

- Bien que la méthode de MSE avec permutation appliquée avec RF-CART ne soit pas directement une allocation de la variance ni non plus une décomposition de la variance comme définies auparavant, après normalisation cette méthode fournit, avec un mtry de 1, un résultat proche de lmg (et de weifila).

- RF-CART par les propriétés mêmes des arbres permet efficacement de prendre en compte les non-linéarités éventuelles dans les données.

Il paraît donc préférable recommander RF-CART avec un mtry de 1 plutôt que *lmg-Shapley* et de ce point de de vue l'analyse de Grömping (2015) ne paraît pas donner à la méthode des forêts aléatoires la place qu'elle mérite. La comparaison faite plus haut avec le jeu *swiss* 182 montre que l'approche RF-CART avec un mtry=1 permet, que

les variables soient ou non quadratisées, de calculer des importances relatives proches de celle de *lmg-Shapley* appliquées aux données quadratisées. Les simulations réalisées ont confirmé l'intérêt de cette méthode par rapport à l'utilisation de *lmg-Shapley* pour la prise en compte des non linéarités. En fait si la simplicité de calcul et de présentation de la méthode est recherchée, il est plus rapide d'utiliser weifila plutôt que *lmg-Shapley*, mais si au contraire la capacité d'intégrer les possibles non linéarités est recherchée, alors RF-CART avec mtry=1 est plus pertinent. Dans tous les cas de figure *lmg-Shapley* n'a ni les avantages de la simplicité de *weifila* ni la capacité de traiter les non-linéarités et interactions comme avec les forêts aléatoires. Enfin le package VSURF permet une sélection des prédicteurs en vue de l'interprétation ou de la prédiction.

Donc nous concluons à plusieurs choix possibles mais aucun ne recommande *lmg-Shapley*.

4.8 Exemple et synthèse.

Avant d'aborder le chapitre 5 qui portera sur les réseaux bayésiens et les approches structurelles et causales, il paraît utile de présenter un exemple de données réelles permettant une vue d'ensemble des résultats obtenus dans un cas concret avec les méthodes présentées aux chapitres 3 et 4, sachant que les données de cet exemple réel seront aussi utilisées dans le chapitre 5 afin d'avoir une vue complète de toutes les approches y compris structurelles. Les données utilisées sont les données *credit* sont accessibles via un lien précisé en annexe 1. Elles comportent 499 observations et 10 variables, dont une variable à prédire et neuf prédicteurs. Ces données sont issues d'une enquête portant sur l'intention de recourir à des services financiers en fonction de réponses relatives aux usages et attitudes par rapport aux vérifications d'assurances crédit et d'acceptations de contrôle par les établissements de notations, pratique très développée aux Etats-Unis. La variable à prédire est : Likelihood to Use. Les 9 prédicteurs sont :

- Business Verification
- Social Verification
- Third Party Verification
- Rating System
- Available Insurance
- Detailed Credit Information
- Report for yourself
- Report for Others
- Credible Third Party

Les résultats de l'ensemble des méthodes suivantes sont présentés dans le graphique ci-après :

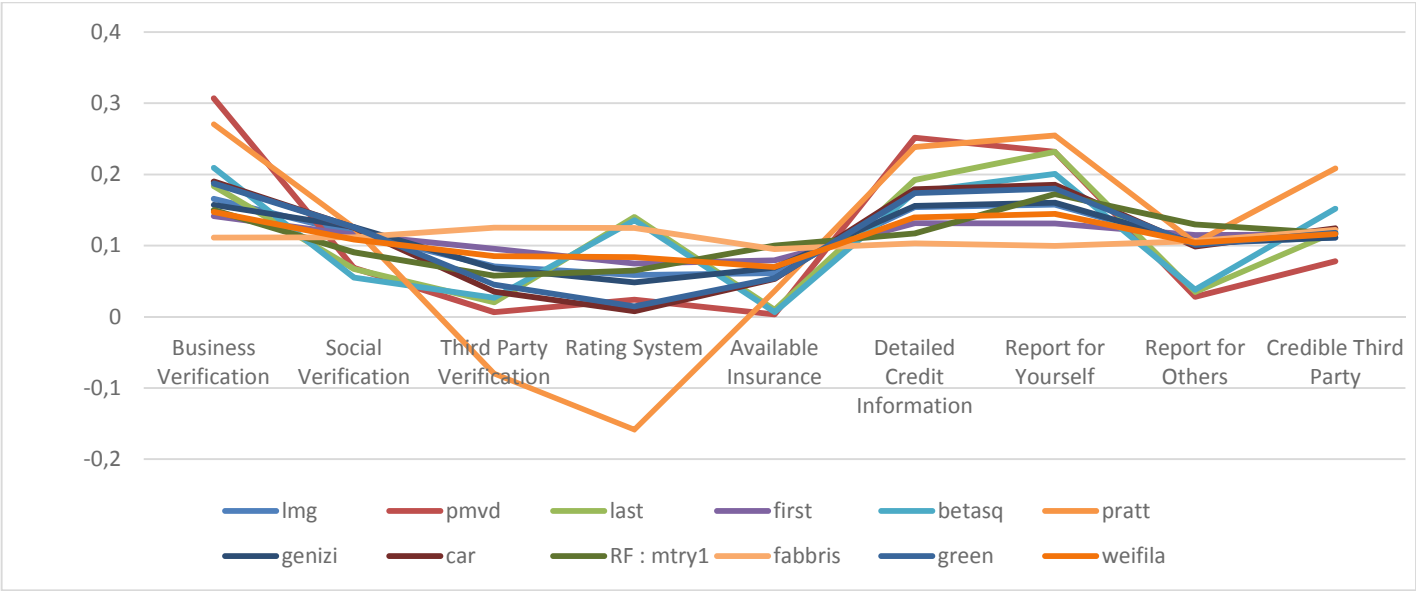


Figure 4.8.1. Comparaisons *lmg-Shapley, pmvd, last, first, betasq, pratt, genizi, car, RF mtry=1, fabbris, green, weifila.* Données *credit*.

Les différentes méthodes présentent des caractéristiques différentes de dispersion dans les valeurs allouées aux neuf prédicteurs. Ceci peut être représenté par leur écart absolu moyen défini comme $\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \bar{x})$.

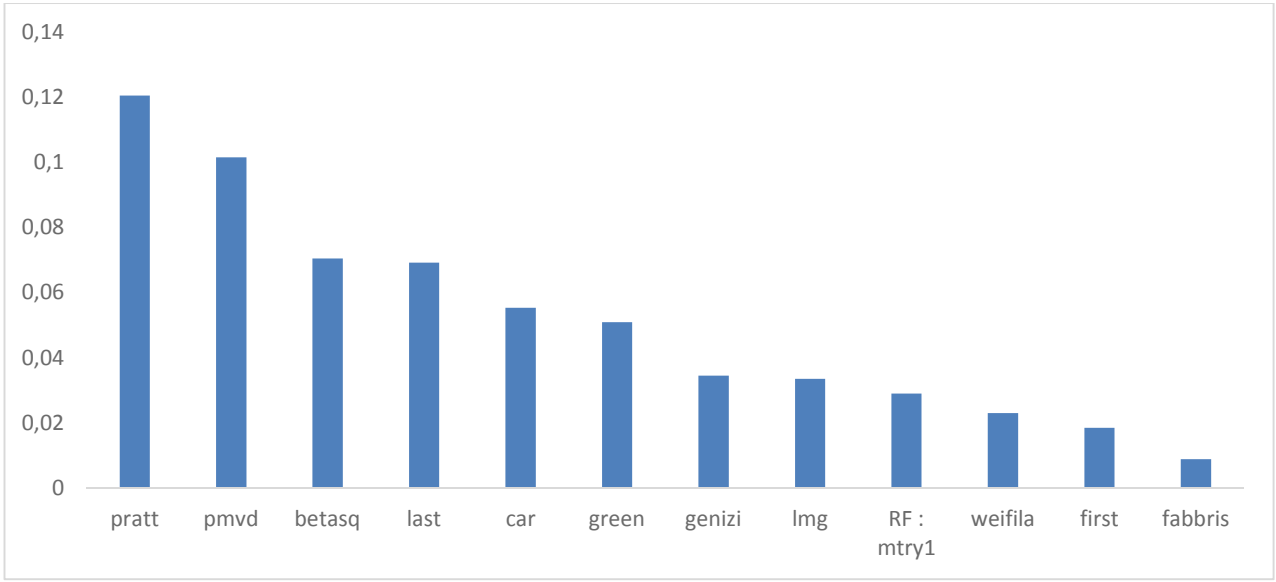


Figure 4.8.2. Ecart absolu moyen des valeurs d'importance normalisées. Données *credit.*

Afin de rendre les résultats du tableau 4.8.1 plus lisibles nous allons présenter séparément les résultats des 12 méthodes en quatre groupes distincts, par ordre décroissant d'écart moyen ;

Groupe 1 : *pratt, pmvd, betasq, last.*

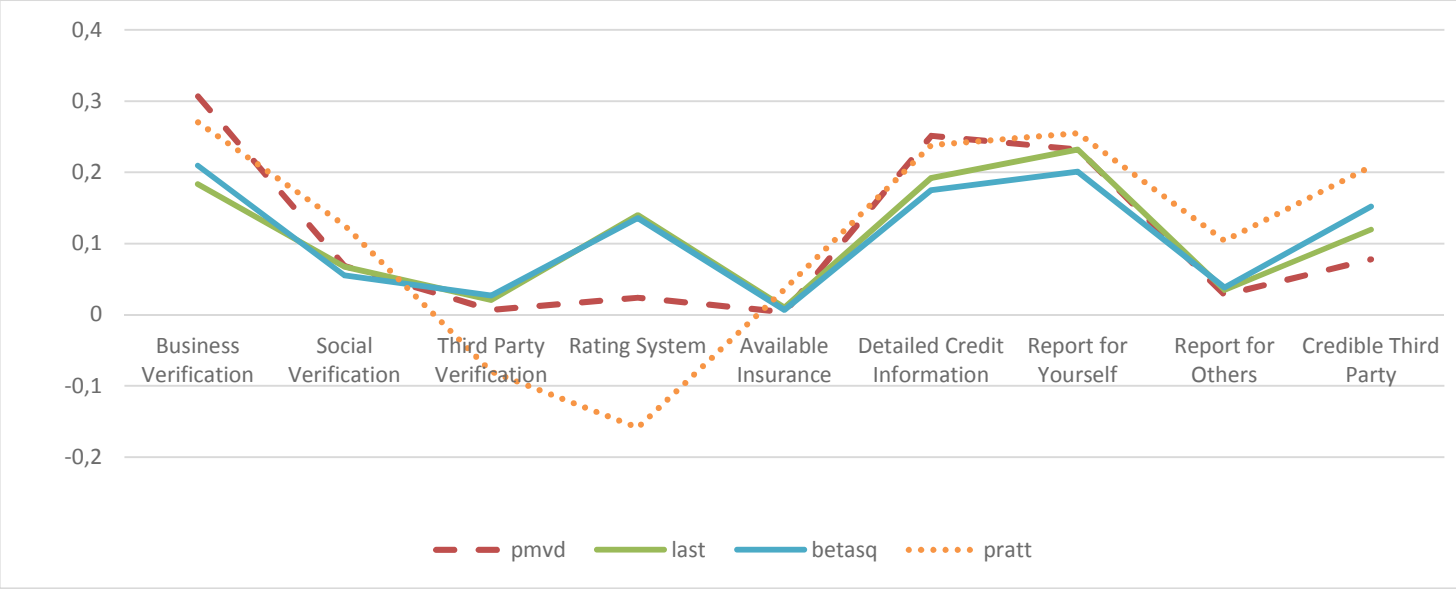


Figure 4.8.3. Comparaisons *pmvd, last, betasq, pratt*. Données *credit.*

Groupe 2 : *car, green.*

.

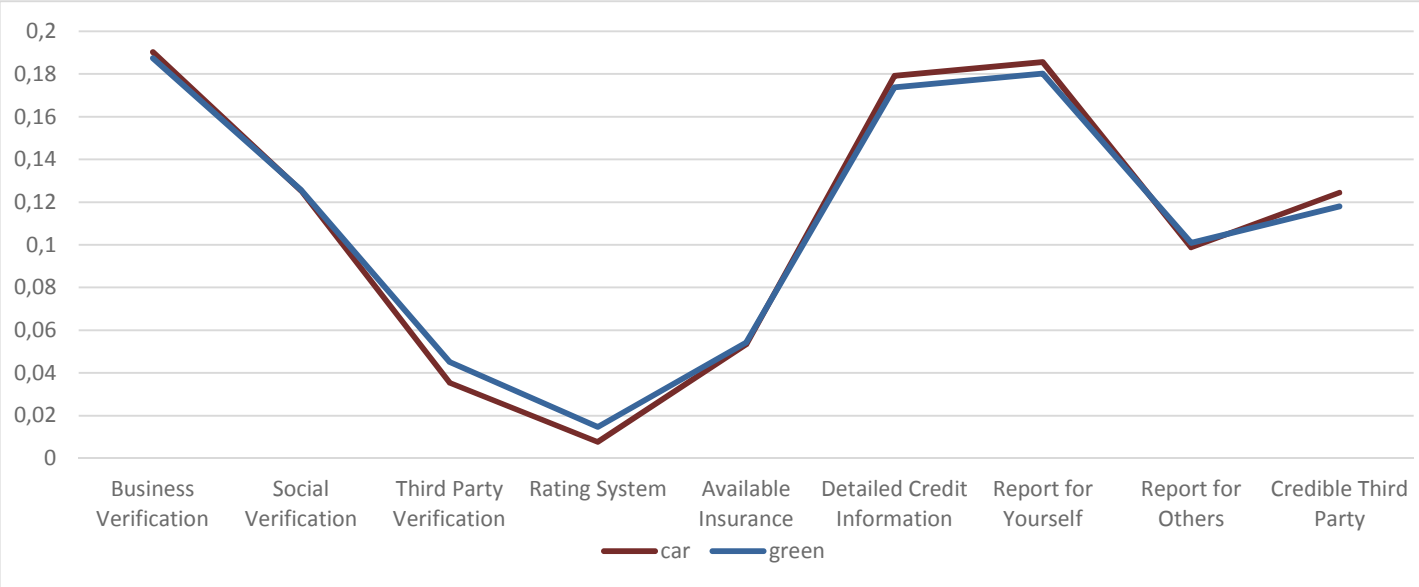


Figure 4.8.4. Comparaison *car, green*. Données *credit.*

Groupe 3 : *genizi (johnson), lmg-Shapley, RF mtry=1, weifila.*

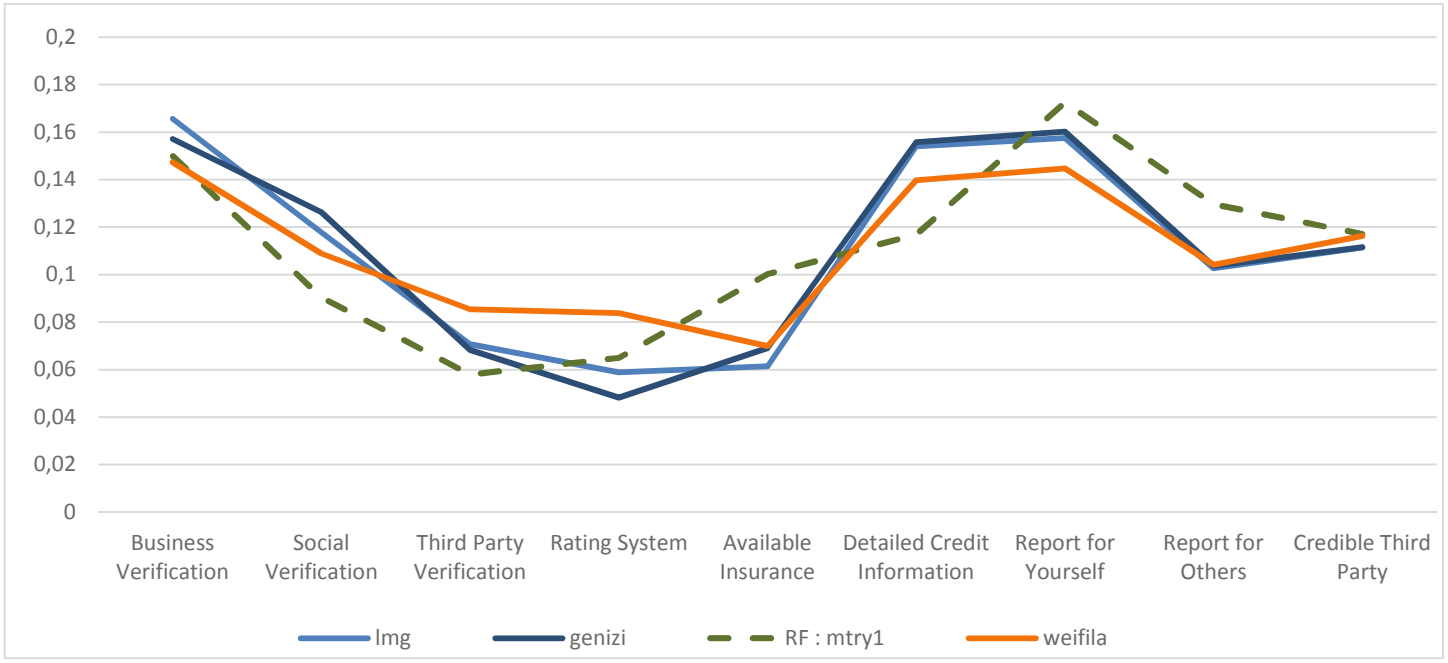


Figure 4.8.5. Comparaisons *lmg-Shapley, genizi-johnson, RF mtry=1, weifila*. Données *credit.*

Groupe 4 : *first, fabbris*

.

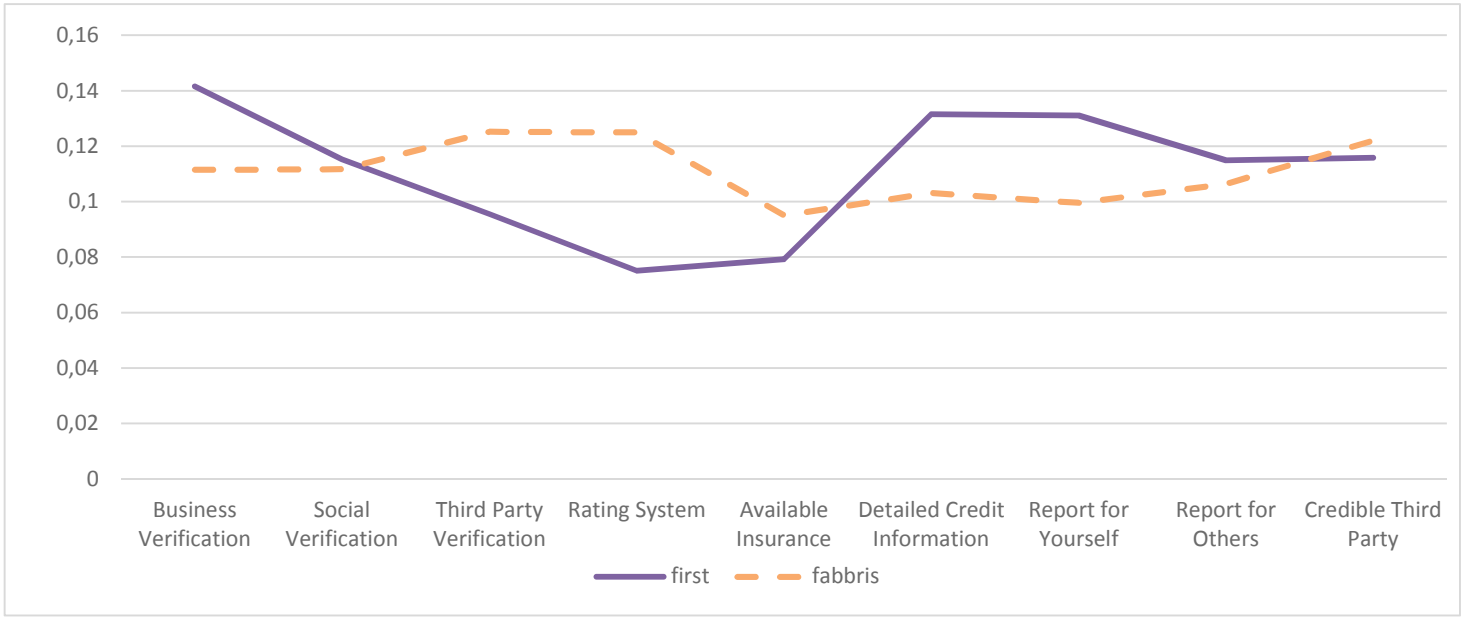


Figure 4.8.6. Comparaison *first, fabbris.* Données *credit*

Enfin il est proposé ci-après un tableau résumant quelques propriétés des différentes méthodes, sachant que ces catégorisations doivent être considérées comme indicatives. Elles permettent de comparer les caractéristiques observées dans le cadre de notre recherche.

Méthode	Dispersion des résultats	Stabilité	Temps de calcul	Complexité
betasq	****	*	*	*
car	***	**	**	***
fabbris	*	****	**	***
first	*	****	*	*
green	***	**	**	***
johnson-genizi	**	***	**	***
last	****	*	*	**
lmg	**	***	****	****
pmvd	****	*	******	*****
pratt	****	*	*	**
RF : mtry1	**	***	***	*****
weifila	**	***	*	**

Chapitre 5 : Vers des analyses causales ?

L'utilisation des réseaux bayésiens dans l'industrie des études de marché a été mise en avant comme une solution nouvelle pour réaliser les analyses des leviers, présentant l'avantage de d'identifier des relations causales et de réaliser efficacement la quantification de l'importance des prédicteurs.

La question de la causalité en statistique est un sujet en soi qui déborde le cadre de cette recherche. Kenett et Salini (2012) ont étudié cette question dans le cadre des études par enquêtes. Dans la pratique des modèles conceptuels existent souvent *a priori* quant à la manière dont se forment par exemple la satisfaction ou l'attachement à une marque. Mais il est aussi intéressant de voir en quoi les réseaux bayésiens peuvent être appliqués avec des données d'études de marché et comment elles peuvent être mises en œuvre pour proposer au praticien une structure de modèle et des méthodes de quantification de l'importance des prédicteurs et contribuer ainsi à la recherche ou à la justification de modèles conceptuels. Après avoir rappelé quelques notions relatives aux réseaux bayésiens et avoir souligné leurs implications, plusieurs exemples d'utilisation dans le domaine de cette recherche seront détaillés. Une analyse des avantages et inconvénients de ces approches sera effectuée et des recommandations seront finalement proposées.

5.1 Réseaux bayésiens

Les modèles graphiques probabilistes et en particulier les réseaux bayésiens, initiés par Judea Pearl dans les années 1980 ont été progressivement utilisés dans de nombreux domaines en raison de leur utilisation commode pour la représentation de connaissances incertaines et la prise en compte de données incomplètes, mais aussi par la possibilité d'incorporer des connaissances *a priori* et des jugements d'experts.

Les principes théoriques et les options d'utilisation des réseaux bayésiens font l'objet d'une abondante littérature et constituent un sujet complexe dont l'étude détaillée dépasse le cadre de cette recherche. Citons en particulier deux ouvrages de référence mentionnés dans la bibliographie: Naïm, P. et al. (2004) et Pearl, J. (2009). De façon synthétique les réseaux bayésiens reposent sur les concepts de théorie des probabilités et de théorie des graphes.

L'utilisation des réseaux bayésiens s'est largement développée depuis 1990 pour des applications professionnelles, facilitées par la mise à disposition de logiciels gratuits ou commerciaux. Ces logiciels permettent d'appliquer des méthodes d'apprentissage des paramètres d'un réseau bayésien à des jeux de données avec ou sans incorporation de connaissance d'un expert.

Cette approche est naturellement attractive dans le cas des études de marchés et sociales d'autant que les outils de visualisation disponibles permettent une communication attractive aux clients et une interactivité forte dans le

processus de définition même du modèle, ainsi que la possibilité de calculs rapides, voire en « temps réel » permettant par exemple de tester immédiatement une variante de modèle et de calculer les impacts correspondants.

Aussi un expert pourra formaliser sa connaissance par des jugements sur certaines valeurs de paramètres, mais aussi sous forme de relations causales qui peuvent être prises en compte dans la modélisation. De l'autre côté, à partir des données pourront être mises en évidence des liaisons entre les différentes variables (indépendances mutuelles, ou conditionnelles). Les outils permettent l'apprentissage supervisé ou non, ainsi que des possibilités de segmentation tant des variables que des observations.

Les applications pratiques sont fondées sur l'utilisation de graphes dirigés sans circuits (DAG pour Directed Acyclic Graphs). Les algorithmes de découverte des arbres font aussi intervenir dans le processus des graphes non dirigés ou des graphes mixtes mais c'est véritablement au final l'utilisation des DAG qui constitue pour le secteur des études de marchés une méthode opérationnelle d'évaluation de l'impact des prédicteurs.

Afin de mettre en lumière les implications dans les interprétations concrètes, quelques définitions, propriétés et résultats méritent d'être rappelés. Ces éléments sont présentés ci-après en s'appuyant sur les travaux de Naïm et al. (2004) et Leray (2006).

Définition : Réseau Bayésien

Un réseau bayésien $B = (G, \theta)$ *est défini par :*

- $G = (X, E)$ *, graphe dirigé sans circuit dont les sommets sont associés à un ensemble de variables aléatoires* $X = \{X_1, ...X_n\}$ *,*

$\theta = \{P(X_i) | Pa(X_i)\}$ *ensemble des probabilités de chaque nœud* X_i *conditionnellement à l'état de ses parents* $Pa(X_i)$ *dans* G *. (Cf. Leray, P. (2006)).*

Pearl et al. ont montré que les réseaux bayésiens permettent de représenter de manière compacte la distribution de probabilités conjointes sur l'ensemble des variables :

$$P(X_{1,}, X_2, ...X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$$

Cette dernière factorisation a rendu possible le développement d'algorithmes d'inférence efficaces qui ont permis aux réseaux bayésiens d'être des outils attractifs pour le traitement de problèmes de raisonnement ou de diagnostics réalisés sur la base de données qu'elles soient complètes ou non.

Comme un réseau bayésien est constitué à la fois d'un graphe et d'un ensemble de probabilités conditionnelles, l'apprentissage comporte deux parties : l'apprentissage de la structure, qui consiste à rechercher le meilleur graphe représentant le problème étudié, et l'apprentissage des paramètres qui se fait en supposant que la structure du réseau a été déterminée. L'apprentissage peut se faire grâce à des données complètes ou non et en utilisant des approches variées y compris bayésiennes et aussi reposer en tout ou partie sur des connaissances d'experts prises en compte dans l'apprentissage.

En ce qui concerne l'apprentissage des paramètres et dans le cas de données complètes, qui est le cas que nous considérons dans cette recherche, la méthode la plus simple et la plus utilisée consiste à estimer la probabilité d'un évènement par la fréquence observée dans les données. C'est en fait une méthode de maximum de la vraisemblance.

$$P(X_i = x_k \mid pa(X_i) = x_j) = \hat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}}$$

Mais il est aussi possible d'utiliser un apprentissage bayésien en utilisant des *a priori* sur les paramètres et de rechercher les paramètres les plus probables avec l'approche de maximum *a posteriori* (MAP) ou encore de calculer l'espérance *a posteriori* (EAP). (Cf. Naïm et al. (2004)). Cette estimation bayésienne est disponible par exemple dans le package R bnlearn en permettant de déterminer un poids de la priore par rapport aux données (cf. Denis, Scutari (2014)). Aussi des méthodes ont été développées pour les situations avec données manquantes et est principalement utilisé l'algorithme EM (Expectation Maximisation) qui peut s'appliquer aussi dans le cadre bayésien.

L'apprentissage de la structure d'un réseau bayésien à partir de données consiste à identifier un graphe correspondant à un modèle d'indépendance associé à une distribution de probabilités conditionnelle qui est dans la pratique observée sur un échantillon d'observations. Deux grandes catégories de méthodes sont utilisées pour trouver les structures de graphe qui représentent le mieux le problème étudié et les données : les premières utilisent des algorithmes dits « par contrainte » et sont fondées sur la recherche d'indépendances conditionnelles et les secondes utilisent des algorithmes recherchant la maximisation d'un score. A côté de ces deux grandes catégories il existe aussi des méthodes hybrides combinant des éléments des deux catégories citées plus haut, ainsi que des méthodes utilisant des graphes partiellement dirigés. Notons que la taille de l'espace des graphes DAG est trop grande pour permettre une recherche exhaustive car le nombre de structures possibles est super exponentiel comme démontré par Robinson (1977). Le nombre de structure avec n nœuds est ainsi donné par la formule récursive suivante :

$NS(n) = 1$ pour n=0 ou 1, puis :

$$NS(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} NS(n-i)$$

Avec 5 nœuds NS(n) est égal à 29281 et avec 10 nœuds : $4,2\ x10^{18}$ (Naïm et al. (2004)).

Différents algorithmes ont donc été développés pour restreindre les recherches et optimiser l'apprentissage de la structure. Dans les cas rencontrés dans les études de marchés avec un nombre de variables de l'ordre de 10 (ou a fortiori nettement plus) la quantification fournie par Robinson montre que l'exploration totale des structures possibles serait irréaliste.

Aussi les réseaux bayésiens ne peuvent pas représenter une distribution quelconque de probabilités conjointes.

Deux hypothèses théoriques doivent être faites :

Hypothèse 1 : Existence d'un réseau bayésien qui soit la représentation du modèle d'indépendance associé à la distribution de probabilité représentée par les données

Hypothèse 2 : Suffisance causale : un ensemble de variables X est suffisant causalement pour une population donnée D si et seulement si dans cette population chaque cause Y commune à plusieurs variables de X appartient aussi a X, ou si Y est constant pour toute la population. (Naïm et al. (2004) page 133). Notons qu'ici la notion de causalité est prise au sens de relation parent enfant dans le graphe.

Une autre notion théorique importante dans ses conséquences pour l'interprétation est celle d'équivalence de Markov. En effet il peut exister plusieurs graphes permettant la représentation d'une loi de probabilité conjointe donnée et donc correspondant à un jeu particulier de données. Deux réseaux bayésiens seront dits équivalents au sens de Markov s'ils représentent les mêmes relations d'indépendances conditionnelles et réseaux équivalents peuvent être regroupés en classes, les classes de Markov. Cette situation est illustrée très simplement ici dans le cas de 3 nœuds (Leray (2006)) :

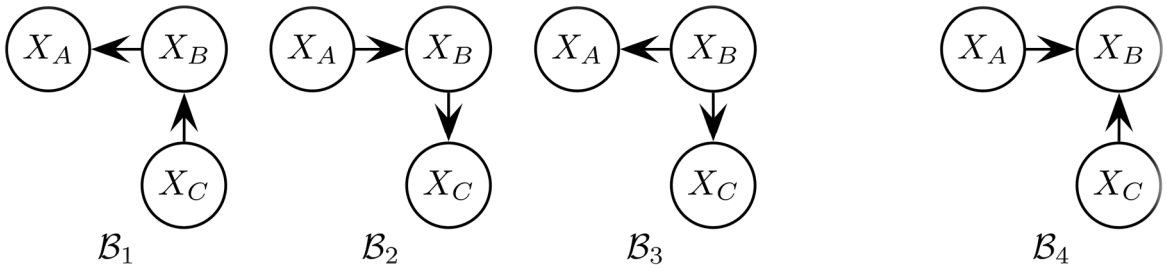


Figure 5.1.1 Equivalence au sens de Markov. B_1 , B_2 et B_3 sont équivalents.

Le graphe B_4 est un exemple de ce qui sera appelé par la suite une v-structure (deux arcs entrants vers le nœud X_B). Plus formellement, voici la définition d'une v-structure (cf. Denis, Scutari (2014)).

Définition : v-structure :

Une connexion convergente $X_i \rightarrow X_k \leftarrow X_j$ *est une v-structure si aucun arc ne connecte* X_i *et* X_j.

Les classes de Markov peuvent également être illustrées par le schéma ci-dessous qui présente les différentes classes d'équivalence de Markov associées à 3 variables : il y a 11 classes de Markov pour l'ensemble des 25 graphes DAG possibles.

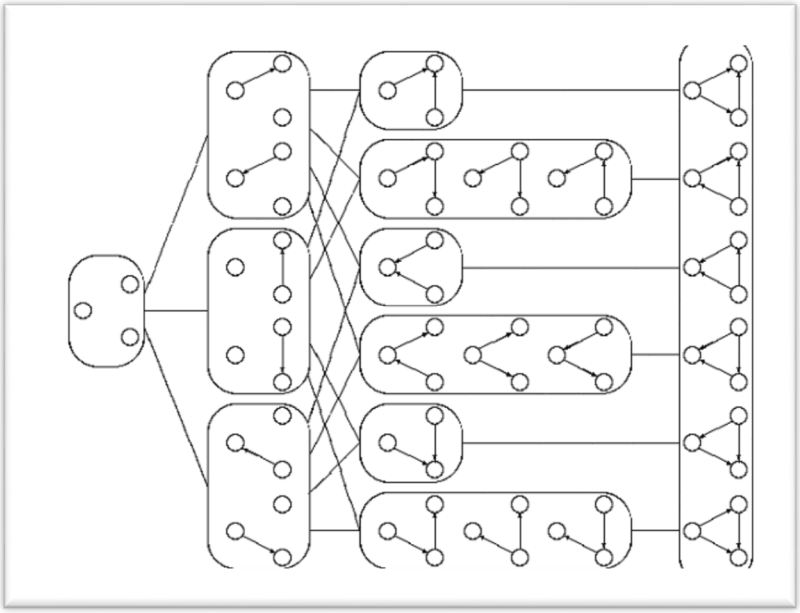


Figure 5.1.2 Classes de Markov associées à 3 variables

Verma et Pearl (1991) ont étudié les classes de Markov et ont démontré que chaque classe d'équivalence peut être caractérisée par un graphe sans circuit partiellement dirigé (PDAG : Partially Directed Acyclic Graph). En effet tous les DAG d'une même classe ont un squelette identique (c'est-à-dire un même graphe non dirigé) quand tous les arcs sont représentés par des liaisons non orientées comme dans les graphes dits de Markov) et possèdent nécessairement les mêmes v-structures.

Une classe peut dont être représentée par le graphe PDAG précité pour lesquels les arcs réversibles (c'est-à-dire qui ne sont pas dans des V structures) ou ceux dont l'inversion n'introduit pas une nouvelle v- structure sont remplacés par des arrêtes non orientées.

Cette propriété théorique d'existence de classes d'équivalence conduit à renforcer la recommandation de prudence quant à la tentation de justification causale des réseaux bayésiens dans les études de marchés : plusieurs structures de DAG peuvent être associés à une même table de probabilités conjointes. Les v-structures sont nécessairement

identiques à l'intérieur d'une même classe de Markov mais certains autres arcs peuvent être inversés et il ne paraît donc pas justifié de les présenter comme tous représentant potentiellement une relation causale.

Notons en outre que dans le cas d'utilisation des réseaux avec bootstrap, c'est-à-dire en ne conservant que des arcs « moyens» entre deux nœuds donnés en fonction d'un critère de sélection (par exemple un arc sera gardé dans le graphe de résultat final si l'apparition de cet arc respecte un seuil donné, comme apparaître dans 50 % au moins des graphes générés sur les différents échantillons du bootstrap), il est théoriquement envisageable que le réseau « moyen » retenu ne corresponde en fait à aucun réseau généré lors du bootstrap.

Le raccourci qui consiste à dire que le réseau bayésien retenu « révèle » une structure causale est donc une simplification excessive. Il paraît préférable de considérer dans le cas général que ce soit avec des méthodes d'indépendances conditionnelles ou des méthodes de score, que les réseaux sont un outil d'aide à la décision du praticien d'études de marché pour décider d'une structure de modèle mais d'éviter de les présenter formellement comme révélant purement et simplement la structure causale.

Les réseaux sont en général définis en termes d'indépendance conditionnelle et de propriétés probabilistes, sans utiliser le fait que les arcs pourraient ou non représenter des relations de cause à effet. L'existence de classes d'équivalence de Markov contenant des DAG avec des arcs de sens opposés entre certains couples de nœuds indique que les arcs ne peuvent être systématiquement en eux-mêmes une indication de causalité. Comme le fait remarquer Scutari (2014), d'un point de vue intuitif un « bon » réseau bayésien devrait pouvoir d'une certaine façon représenter la structure causale du phénomène observé. Cependant la question de la causalité doit être maniée avec une grande prudence à la fois pour des raisons de portée générale mais aussi pour des raisons de nature plus spécifiques liées au contexte des études de marché.

En ce qui concerne l'interprétation causale de façon générale, Pearl (2009) cité par Scutari (2014) indique dans son livre sur la causalité : *« Il semblerait que si les assertions sur les indépendances conditionnelles proviennent de relations causales admises, alors enregistrer et représenter ces relations directement serait une façon plus naturelle et plus fiable d'exprimer ce que nous savons ou imaginons à propos du monde. C'est la philosophie qui se trouve derrière les Réseaux Bayésiens causaux ».* L'apprentissage de relations causales en général représente donc une question difficile et nécessitant des précautions. Dans le cas spécifique des études de marché effectuées sur la base de questions posées à des répondants à un instant donné, l'interprétation des relations de causalité entre les phénomènes observés via ces questions peut être légitimement fondée sur une théorie sous-jacente, mais en évitant de trop tenter de justifier des relations causales simplement à partir des données.

Bühlmann (2013) a travaillé spécifiquement sur l'identification des DAG au sein des classes de Markov et sur l'inférence causale dans le cas de données en grande dimension issues de la biologie. Prenant en considération le fait que les arcs dirigés ne sont pas toujours identifiables Bühlmann examine la possibilité de déterminer des bornes aux « causal effects ». Il relève le fait que dans les cas où le réseau est suffisamment « sparse » c'est-à-dire s'il y a suffisamment d'arcs absents entre les différents nœuds, le graphe peut être identifiable.

Ceci est illustré sur la figure suivant où le DAG est en fait unique dans sa classe d'équivalence de Markov.

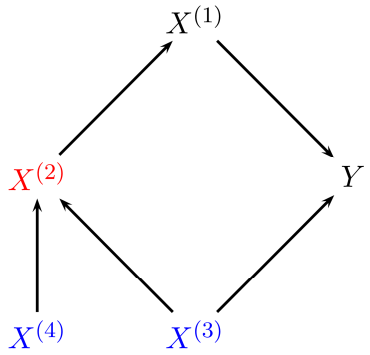


Figure 5.1.4 Graphe identifiable unique dans sa classe de Markov. (cf. Bühlmann (2013)

Dans les autres cas, pour estimer des bornes d'effets causaux (« bounds of causal effects ») Bühlmann a recours à la notion d'intervention qui consiste pour estimer l'impact direct et indirect d'une variable à la fixer à une certaine valeur, ce qui permet grâce au DAG obtenu de déterminer une distribution conditionnelle de la variable à prédire.

$$P(Y \left| do(X_j = x)\right))$$

Il définit alors l'effet d'intervention (« intervention effect aussi appelé « causal effect ») au point x_0 comme :

$$\frac{\partial}{\partial x} E\left[Y \left| do(X_j = x)\right.\right]\Big|_{x=x_0}$$

Dans le cas où les prédicteurs et la variable à prédire ont une distribution gaussienne multivariée $E\left[Y \left| do(X_j = x)\right.\right]$ est une fonction linéaire de x et l'effet d'intervention est alors un paramètre :

$$\theta_j \equiv \frac{\partial}{\partial x} E\left[Y \left| do(X_j = x)\right.\right] \qquad (j = 1,...p).$$

θ_j est coefficient de régression de X_j dans la regression linéaire de Y sur $\left\{X_j, X^{(pa(j))}\right\}$ (Bühlmann (2013) et Pearl (2000) qui a recours au « back door criterion »).

Les bornes des effets causaux peuvent alors être estimées sur l'ensemble des DAG d'une classe de Markov. Cette méthode a été mise en œuvre avec le package R pcalg et décrite par Bühlmann et al (2012) et est appelée la méthode IDA. Nous présenterons plus loin des exemples de résultats obtenus en utilisant IDA dans des cas simples, sachant que cette méthode a été développée dans le cadre des biostatistiques avec un très grand nombre de variables.

Ces éléments théoriques liés aux réseaux bayésiens ont une grande importance au regard de l'utilisation qui en est faite dans le cas des études de marchés. La multiplicité des graphes possibles et les précautions nécessaires dans l'interprétation de la causalité devraient être prises en considération dans la manière dont ces méthodes sont utilisées et présentées aux clients quand elles sont appliquées à des résultats d'enquête, ce qui n'est pas toujours le cas soit en raison du manque de familiarité des praticiens avec les concepts théoriques de ces approches soit parce que la communication commerciale exige un message simple.

En ce qui concerne l'apprentissage des structures voici une description synthétique des principales méthodes.

5.1.1 Méthodes par contraintes.

Ces méthodes ont été développées par plusieurs équipes :

- Spirtes, Glymour et Scheines ont développé plusieurs algorithmes (SGS en référence aux trois auteurs, PC en référence à Peter et Clark, CI, FCI pour « Fast Causal Inference »)
- Pearl et Verma ont développé IC (pour Inductive Causation) et IC*, qui permet de détecter éventuellement si il peut exister des variables latentes
- Cheng et al. ont plus récemment développé BN-PC, pour Belief Net Power Constructor
- PMMS pour Polynomial Max-Min Skeleton, par Brown et al.

Ces méthodes nécessitent la détermination des relations d'indépendance conditionnelles entre les variables et donc l'utilisation de tests statistiques d'indépendance.

Les tests peuvent être soit le test du χ^2 soit le rapport de vraisemblance G^2 ainsi défini :

Rapport de Vraisemblance

Soient deux variables discrètes X_A *et* X_B *de taille respective* r_A *et* r_B N_{ab} *le nombre d'occurrences de* $\{X_A = x_a, X_B = x_b\}$. *Soit* $N_{a.}$ *le nombre d'occurrences de* $\{X_A = x_a\}$ *et* $N_{b.}$ *le nombre d'occurrences de* $\{X_B = x_b\}$

alors

$$G^2 = 2\sum_{a=1}^{r_A}\sum_{b=1}^{r_B} O_{ab} \ln(\frac{O_{ab}}{T_{ab}}) = 2\sum_{a=1}^{r_A}\sum_{b=1}^{r_B} N_{ab} \ln(\frac{N_{ab}N}{N_{a.}N_{b.}})$$

G^2 suit aussi asymptotiquement une loi du χ^2 de degré de liberté, $df = (r_A - 1)(r_B - 1)$ sous l'hypothèse que les variables soient indépendantes et identiquement distribuées.

Ces tests sont utilisés avec des adaptations pour tenir compte des situations où l'indépendance est testée avec plusieurs variables conditionnelles et où le nombre d'observations peut être limité. Plus de détails sont fournis dans Naïm et al. (2004)

Dans le cas par exemple de l'algorithme PC, le principe consiste à partir d'un graphe non-dirigé où tous les arcs sont présents, à éliminer les liaisons entre variables en fonction des tests d'indépendance, en supprimant les arcs liant des variables en cas d'indépendance d'ordre 0, 1, etc. puis à identifier les v-structures et finalement à compléter le graphe avec des arrêtes respectant la contrainte d'un DAG. Ces différents algorithmes ont fait l'objet au fur et à mesure de perfectionnement en vue de diminuer le nombre de tests d'indépendances à réaliser. (Leray (2006)).

5.1.2 Méthodes d'optimisation d'un score.

A la différence des approches fondées sur les tests d'indépendances conditionnelles qui procèdent par étapes en analysant successivement des arcs particuliers, les approches fondées sur des scores de réseaux prennent en compte chaque DAG ou réseau partiellement dirigé dans son ensemble. Elles vont soit chercher une structure qui optimise un certain score soit chercher un jeu de meilleures structures et en combiner les résultats.

Afin de pouvoir être efficacement mis en pratique ces scores doivent être décomposables localement au sens où ils peuvent s'exprimer comme la somme de scores locaux et finalement le score global est la somme de scores calculés au niveau de chaque nœud du réseau ou de parties du réseau. Comme montré précédemment le nombre de réseaux possibles est super exponentiel par rapport au nombre de nœuds donc les algorithmes opèrent sur un espace réduit ou effectuent une recherche gloutonne (greedy algorithm) dans cet espace, c'est-à-dire en choisissant étape par étape un optimum local en tentant in fine d'atteindre un optimum global.

Les scores utilisés appliquent le principe de parcimonie souvent référencé comme le principe du rasoir d'Occam c'est-à-dire en recherchant un modèle le plus simple possible tout en étant ajusté aux données. Ces scores peuvent

ainsi être composés de deux termes : un terme de vraisemblance et un terme de complexité du modèle. La complexité du modèle peut être quantifiée par une quantité associée au nombre de paramètres indispensables à la représentation du réseau. Ce nombre de paramètres peut être calculé comme suit :

Soit X_i *un nœud du réseau de taille* r_i *et* $pa(X_i)$ *ses parents, pour décrire* $P(X_i \mid pa(X_i))$ *il est nécessaire d'avoir* $Dim(X_i, B)$ *paramètres avec :*

$$Dim(X_i, B) = (r_i - 1) \prod_{X_j \in pa(X_i)} r_j = (r_i - 1) q_i$$

et le nombre de paramètres requis pour l'ensemble du réseau est :

$$Dim(B) = \sum_{i=1}^{n} Dim(X_i, B) = \sum_{i=1}^{n} (r_i - 1) q_i$$

Il existe de nombreux scores, citons :

- Entropie conditionnelle
- Critères AIC et BIC
- Longueur de description minimale (MDL pour Minimum Description Length)
- BD: Bayesian Dirichlet
- BDe: Bayesian Dirichlet Equivalent
- BDγ : Bayesian Dirichlet généralisé

Ces scores ou leur logarithme sont des scores décomposables au sens où (Leray (2006)) :

$$Score(B, D) = cste + \sum_{i=1}^{n} score(X_i, pa(X_i))$$

Avec des scores décomposables il est possible de rechercher des optimisations de parties du réseau pour optimiser la recherche :

- utilisation de l'arbre de recouvrement maximal (Maximum Weight Spanning Tree) : arbre incluant tous les nœuds en optimisant le score
- ordonnancement des nœuds en ajoutant un ordre

- recherche gloutonne dans l'ensemble des graphes.

L'analyse détaillée de ces techniques dépasse le cadre de cette recherche mais des compléments peuvent être trouvés dans les ouvrages de référence sur le sujet.

Certains algorithmes permettent de traiter des variables continues, mais les analyses considérées dans cette recherche portent sur des variables discrètes ou discrétisées lorsque des réseaux bayésiens sont mis en œuvre.

Malgré la relative complexité théorique associée aux réseaux bayésiens, la possibilité de prendre en compte des avis d'experts et les capacités de visualisation ont contribué à susciter un grand intérêt dans le secteur des études de marchés. Certains outils sont à la disposition des utilisateurs depuis au moins une dizaine d'années comme Bayes Net Toolbox, bibliothèque open-source de fonctions Matlab, BayesiaLab proposée par la société française Bayesia ou encore HuginExpert (société Hugin) et Netica de la société Norsys.

Aussi il existe différents packages dans R et un site spécialisé www.bnlearn.com qui propose également des publications associées. Enfin il existe naturellement des applications propriétaires développées par certaines sociétés.

Le site http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html liste plus de 60 solutions, donc les options sont aujourd'hui nombreuses.

5.2 Exemples d'application

La variabilité des méthodes s'accompagne de façon assez logique d'une variabilité des résultats comme montré par Kenett et Salini (2012). La figure ci-après montre les deux réseaux obtenus sur le même jeu de données selon qu'est utilisé une construction du réseau par indépendances conditionnelles ou une méthode de score.

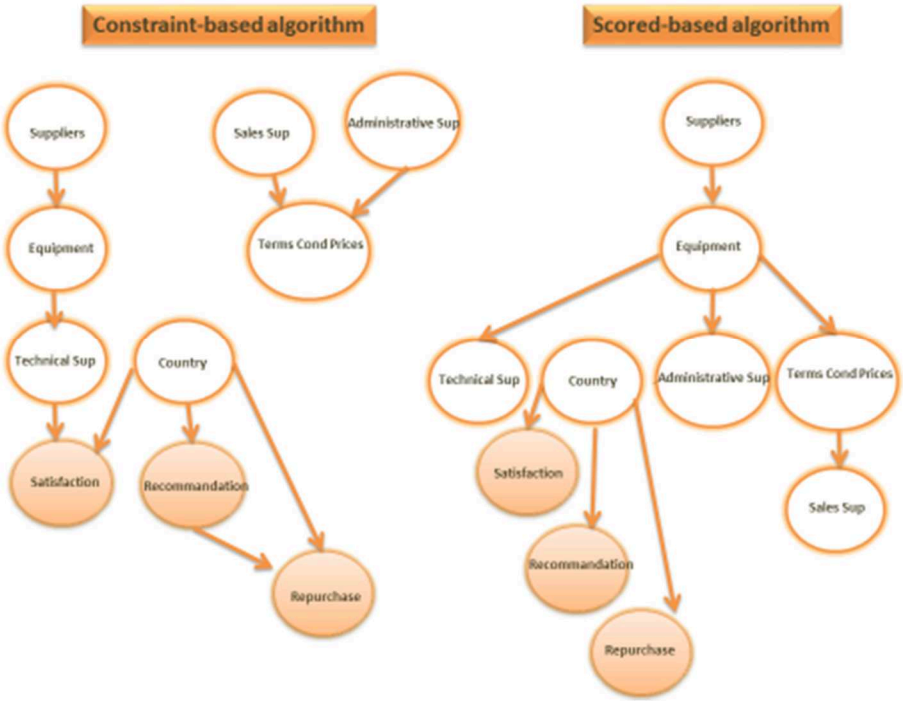


Figure 5.2.1. Comparaison des résultats appliqués à un jeu de données. (d'après Kenett et Salini (2012)).

Cet exemple simple illustre la variabilité potentielle des réseaux obtenus en fonction du choix des méthodes et montre bien l'importance d'une analyse d'expert pour valider la structure. Autant le modèle obtenu ci-dessus avec une méthode de score est assez interprétable en considérant que la satisfaction conditionne directement la recommandation et l'intention d'achat, autant le modèle par contrainte est plus difficile à interpréter.

Aussi il aurait été tout à fait possible de considérer une variable latente associée à « satisfaction », « recommandation » et « intention d'achat » qui pourrait être appelée « Expérience », et qui elle-même conditionnerait ces trois variables.

De fait la variété des options techniques disponibles et les seuils de sélections qui peuvent être retenus en ce qui concerne la force des liaisons entre nœuds conduit potentiellement à une très grande variété des résultats en ce qui concerne les choix de structure, ce qui conduit à des interprétations multiples quant aux liens entre variables, et aussi dans une certaine d'estimation finale des impacts lors de l'utilisation d'un modèle donné pour simuler des impacts.

Notre recherche a consisté à comparer différents outils d'apprentissage, à examiner leurs conditions de mise en œuvre et à comparer les résultats obtenus avec différences méthodes sur un même jeu de données.

5.2.1 Outils et méthodes

Le package bnlearn

Le package « bnlearn » offre un ensemble varié de méthodes d'apprentissage des réseaux et d'estimation des paramètres. Les méthodes d'apprentissage comportent quatre algorithmes pat contraintes aussi appelés dans la documentation « conditional independance learners » et deux algorithmes par maximisation de score ainsi que deux algorithmes hybrides permettant une optimisation de la recherche de réseaux.

De plus bnlearn fournit quatre algorithmes supplémentaires de découverte des squelettes (graphes non dirigés), types d'arbres qui comme indiqué plus haut peuvent être utilisés dans l'apprentissage avec des méthodes hybrides.

Les méthodes contenues dans bnlearn sont :

Algorithmes par contraintes (Constraint-based algorithms)

- Grow Shrink (gs)
- Incremental Association (iamb)
- Fast Incremental Association (fast.iamb)
- Interleaved Incremental Association (inter.iamb)

Algorithmes par maximisation de score (Score-based Learning algorithms)

- Hill-Climbing (hc)
- Tabu Search (tabu)

Algorithmes Hybrides

- Max-Min Hill-Climbing (mmhc)
- Restricted Maximisation (rsmax2)

Graphes non dirigés

- Max-Min Parents and Children (mmpc)
- Hiton Parents and Children (si.hiton.pc)
- Chow-Liu (chow.liu)
- ARACNE (aracne)

Plusieurs options d'optimisation ainsi que des variantes sont disponibles pour ces différents algorithmes. Ces éléments sont décrits en détail dans la documentation R ainsi que sur le site http://www.bnlearn.com. L'auteur du package bnlearn, Marco Scutari a également publié un ouvrage édité en français (2014). Le simple énoncé des différentes capacités offertes par bnlearn montre la variété des techniques et donc des résultats possibles sur un même jeu de données. Ceci s'accompagne des différentes options offertes dans la construction des réseaux par exemple en imposant que certains arcs soient présents (white list) ou au contraire exclus (black list).

BayesiaLab

BayesiaLab est la solution commercialisée par la société Bayesia (www.bayesia.com). Développé en Java, BayesiaLab permet de traiter les différentes étapes de la modélisation par un réseau bayésien, et est très avancée en matière d'interfaces et d'apprentissage. Elle utilise des algorithmes performants et offre aussi une facilité de chargement des données et de gestion des options de traitement.

Dans le cadre de cette recherche l'attention a été portée sur la méthode proposée par Bayesia dans sa documentation technique pour l'analyse des leviers. Cette approche peut être décrite synthétiquement de la façon suivante :

1. Importation des données
2. Discrétisation et agrégation
3. Réalisation d'un graphe non-supervisé réalisé en ne gardant que les variables prédictrices c'est-à-dire en ayant laissé de côté la variable à prédire
4. Analyse préliminaire du réseau obtenu pour une première interprétation de la structure
5. Identification de facteurs (variable clustering) en utilisant la divergence de Kullback-Leibler. Le nombre de facteurs peut être choisi arbitrairement ou automatiquement décidé par l'application.
6. Interprétation des facteurs en liaison avec les variables associées, il s'agira par exemple d'éléments marketing, comme « familiarité » ou relationnels, comme « accueil en point de vente ».
7. Elaboration d'un modèle probabiliste d'équations structurelles (PSEM : Probabilistic Structural Equation Model). Ceci est réalisé avec les options suivantes et la fixation de règles restreignant les choix de réseaux possibles :
    - Apprentissage supervisé (la variable à prédire est le « target node ») incluant outre les variables de départ aussi les facteurs identifiés au point 5.
    - Interdiction d'avoir des arcs entre les prédicteurs
    - Interdiction d'avoir des arcs orientés des prédicteurs vers les facteurs
    - Interdiction des arcs entre les prédicteurs et la variable à prédire

Il est alors possible d'effectuer une analyse du type IPA (Importance Performance Analysis) ; l'importance étant quantifiée par l'information mutuelle des différents facteurs avec la variable à prédire c'est-à-dire en fonction des observations :

$$IM(X;Y) = \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} \frac{O_{ij}}{N} \log(\frac{O_{ij}N}{E_{ij}})$$

où O_{ij} représente le nombre de fois où la configuration x=i et y=j apparaît dans la base et avec

$E_{ij} = N_{i.}N_{j.} = (\sum_{i=1}^{s_1} O_{ij})(\sum_{j=1}^{s_2} O_{ij})$ et N est le nombre de cas dans la base.

Il est aussi possible de simuler l'impact d'un changement des performances sur les différents prédicteurs sur la variable à prédire en faisant varier les valeurs de ces prédicteurs et en utilisant les paramètres du réseau généré pour calculer des impacts. L'utilisation de BayesiaLab permet donc de choisir un modèle de façon très interactive, en incorporant le choix du nombre de facteurs et éventuellement en choisissant grâce à un jugement d'expert les variables entrant dans le modèle.

Les fonctionnalités de BayesiaLab permettent également de présélectionner certains arcs et d'élaborer des modèles incorporant des causalités décidées *a priori,* ce qui signifie qu'il existe bien d'autres manières de modéliser avec cette solution que l'approche spécifique PSEM générique telle que décrite plus haut.

La contrepartie de ces flexibilités dans le choix du modèle fait que cette approche ne génère pas une quantification d'importance de portée générale comme les méthodes étudiées aux chapitres 2 et 3. Notons également que l'importance utilisée dans l'analyse IPA présentée plus haut peut être comparée à une mesure d'association ou corrélation simple entre un des facteurs identifiés et la variable à prédire, et que les impacts calculés par simulation peuvent être comparés à ceux obtenus en utilisant une modélisation structurelle ou même une régression multiple.

Ces commentaires seront aussi applicables dans l'autre exemple de méthode utilisé qui est présenté ci-après.

IBN

Cette méthode a été décrite par Egner et Hart (Ipsos, 2012) et est fondée sur l'utilisation des probabilités conditionnelles avec le programme PC algorithm (package `pcalg`).

Les étapes principales sont :

1. élimination des arcs en cas d'absence de significativité du test de corrélation bivariée

2. élimination des arcs en cas de non indépendance conditionnelle vis-à-vis d'un nœud dans la couverture de Markov

3. orientation des arcs selon les v-structures avec choix de V fait en fonction du test le plus fort en cas de conflit de direction (bidirectionnalité) et possibilité de double arc si l'algorithme ne détermine pas de direction dans une liaison entre deux nœuds

Pour chaque réseau généré, un calcul des importances est réalisé par simulation et propagation en utilisant la modification des probabilités conditionnelles à chaque pas. Pour les cas où un nombre d'observations est trop limité les probabilités conditionnelles sont calculées en agrégeant les niveaux des variables prédictives en utilisant un CART. Un ré échantillonnage (bootstrap) est réalisé pour générer 500 graphes

A ce stade deux agrégations sont réalisées. Premièrement un maintien des arcs qui apparaissent au-dessus d'un certain seuil d'occurrence. Ce seuil est un paramètre qui peut être choisi. Puis deuxièmement un calcul de la moyenne des scores d'importance obtenus avec chacun des graphes du bootstrap.

5.2.2 Résultats

Plusieurs comparaisons ont été effectuées entre les différentes méthodes :

- différences de résultats entre les 12 algorithmes proposés par bnlearn,
- utilisation de données de référence du package bnlearn (gaussian.test et learning.test) à la fois avec le package bnlearn et avec IBN
- application de bnlearn à des données étudiées antérieurement (UK Data et *swiss* 182) et comparaisons avec IBN.

Données « insurance »

Les données « insurance » sont incluses dans le package bnlearn ont été utilisées pour comparer les graphes obtenus selon les types de méthodes utilisées. Il comprend 27 variables et 20000 observations et est décrit sur le site de bnlearn : (http://www.bnlearn.com/documentation/man/insurance.html).

Les résultats des différents algorithmes sont très variés. Ainsi le nombre d'arcs avec les 4 méthodes par contraintes varie de 28 à 39 arcs et les deux méthodes de maximisation d'un score génèrent 50 arcs. C'est la méthode hybride « Restricted Maximisation » qui génère le moins d'arcs, 20, tandis que l'algorithme Hiton Parents and Children délivre le maximum d'arcs : 416.

Cette utilisation directe et sommaire, c'est-à-dire sans ajouter les différentes options, confirme la variété des résultats et met en évidence que certains algorithmes qui vont « sur-spécifier » le réseau en accroissant le nombre d'arcs.

Quelques graphes illustratifs avec le jeu « insurance » sont présentés ci-dessous :

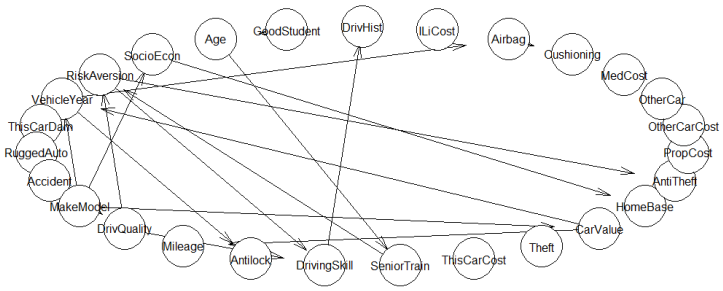


Figure 5.2.2 1 Données insurance. Graphe avec Restricted Maximisation.

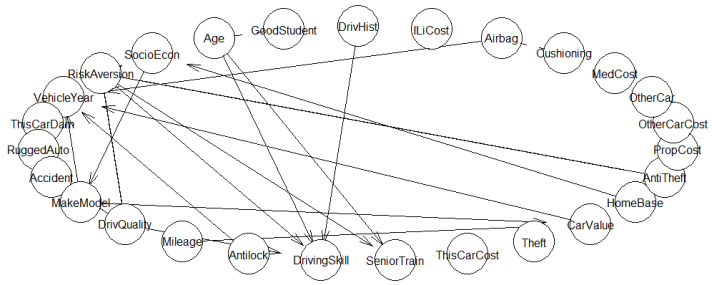


Figure 5.2.2 2 Données insurance. Graphe avec Grow Shrink.

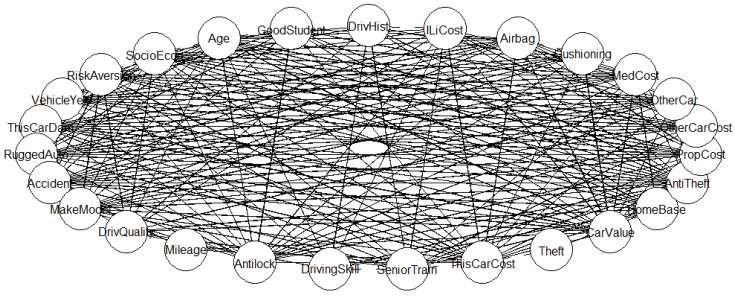


Figure 5.2.2 3. Données insurance. Graphe avec Hiton Parents and Children.

Données « gaussian.test »

Le jeu « gaussian.test » est aussi inclus dans le package bnlearn. Il est composé de 5000 observations et 7 variables. Ici les 4 méthodes par contraintes donnent les mêmes réseaux avec 8 arcs, et les deux méthodes de score et les deux méthodes hybrides donnent toutes quatre également les mêmes réseaux avec 7 arcs. Comme précédemment c'est la méthode Hiton Parents and Children qui donne le plus grand nombre d'arcs : 18, tous les arcs de ce dernier graphe étant bidirectionnels.

Les graphes illustratifs de ces ensembles de résultats sont figurés ci-dessous :

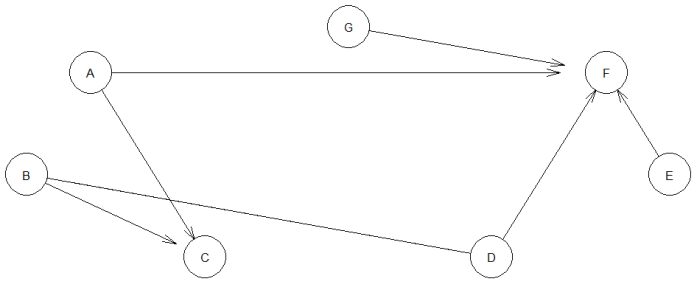


Figure 5.2.2 4. Données gaussian.test. Graphe avec GS.

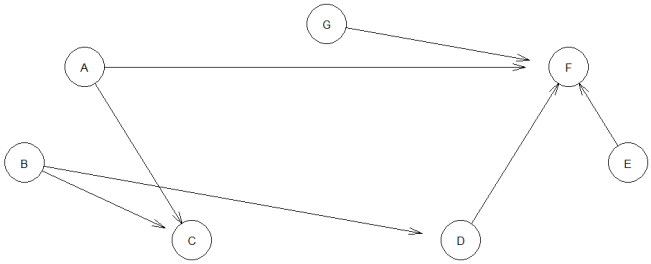


Figure 5.2.2 5. Données gaussian.test. Graphe avec Restricted Maximisation.

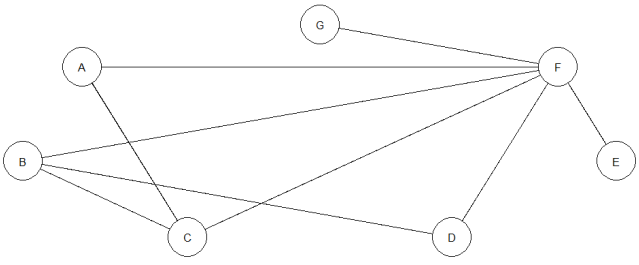


Figure 5.2.2 5. Données gaussian.test. Graphe avec Hiton Parents and Children.

Le réseau obtenu avec IBN est le suivant :

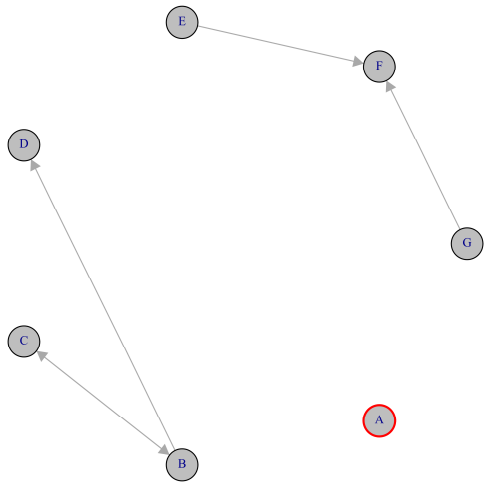


Figure 5.2.2 6. Données gaussian.test. Graphe avec IBN.

L'effet de sélection des arcs est plus prononcé avec IBN en raison de l'existence d'un seuil de sélection avec les bootstrap.

Données « learning.test »

Ces données sont aussi incluses dans le package bnlearn et comportent 5000 observations et 6 variables nommées A, B, C, D, E, F.

Les quatre méthodes par contraintes de bnlearn donnent toutes le même réseau avec 6 arcs avec AB bidirectionnel.

Les deux méthodes de score et les deux méthodes hybrides donnent toutes quatre le même réseau avec 5 arcs, AB étant cette fois orienté seulement de A vers B.

Enfin les 4 derniers algorithmes donnent également le même réseau avec 10 arcs.

Ils sont présentés ci-dessous :

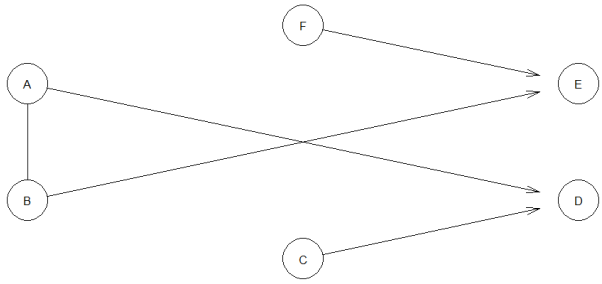


Figure 5.2.2 7. Données learning.test. Graphe avec Grow Shrink.

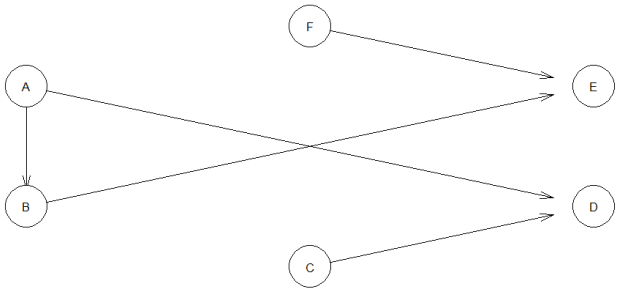


Figure 5.2.2.8. Données learning.test. Graphe avec Restricted Maximisation.

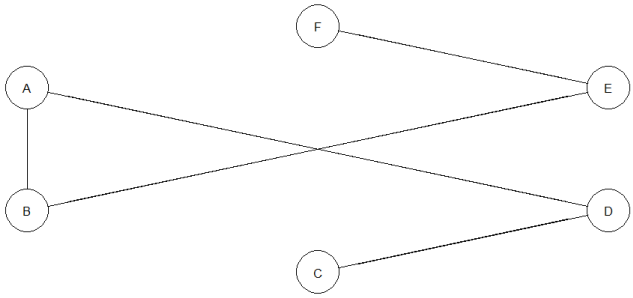


Figure 5.2.2.9 Données learning.test. Graphe avec Hiton PC.

L'utilisation d'IBN délivre un réseau avec 6 arcs bidirectionnels, excluant les variables C et F.

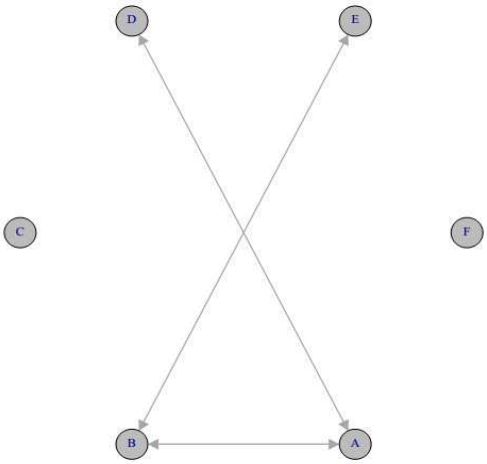


Figure 5.2.2.10. Données learning.test. Graphe avec IBN.

Données UK Data

Les données UK Data ont déjà été présentées. Les quatre méthodes par contraintes donnent des graphes comprenant entre 55 et 57 arcs, résultat proche de ceux obtenus avec les méthodes hybrides qui génèrent des graphes de 54 et 56 arcs. Le graphe le plus dense est encore une fois obtenu avec Hiton Parents and Children avec 210 arcs. Le graphe le plus parcimonieux est obtenu avec l'algorithme Chow-Liu 28 nœuds.

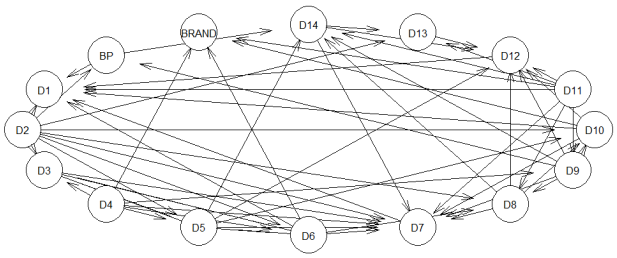


Figure 5.5.2 11. Données UK Data. Graphe avec GS.

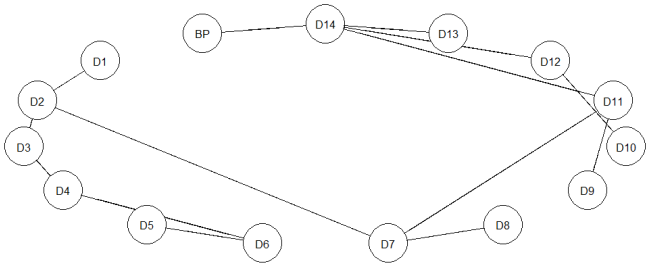


Figure 5.5.2.12. Données UK Data. Graphe avec Chow-Liu.

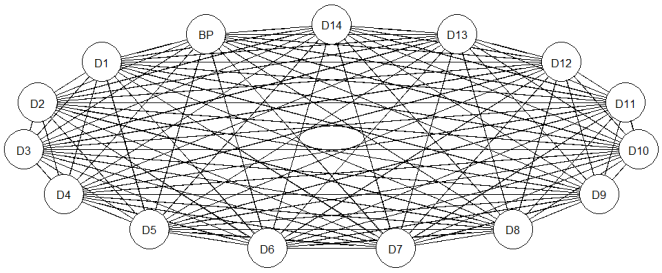


Figure 5.5.2.13. Données *UK Data*. Graphe avec Hiton Parents and Children.

Nous avons également calculé les importances allouées aux prédicteurs en utilisant bnlearn mais aussi `pcalg` avec la méthode IDA décrite ci-dessus. Voici le graphe obtenu avec bnlearn Hill Climbing :

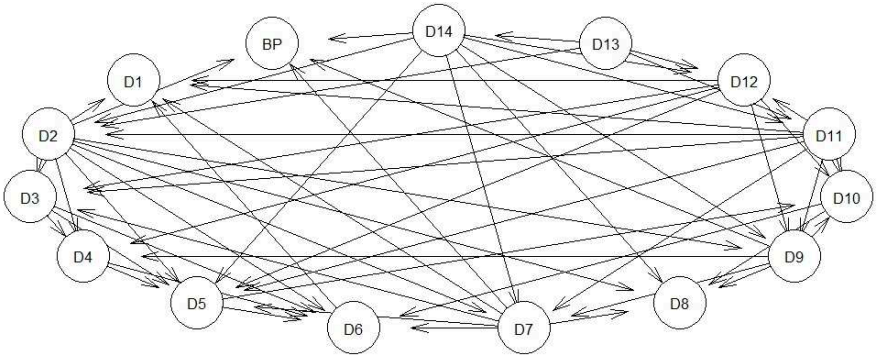


Figure 5.5.2.14. Graphe bnlearn. Hill Climbing. UK Data.

Dans ce cas la variable d'intérêt BP a quatre parents dans le graphe : D2, D7, D9 et D14. L'utilisation de la fonction `predict` de `bnlearn` appliquée à la prévision de la variable BP en utilisant le réseau bayésien ajusté aux données (bn.fit) conduit à simplement utiliser la régression de la variable d'intérêt BP sur ces 4 prédicteurs.

Nous avons à titre de confirmation effectué à la fois une simulation via bnlearn et une régression directe et trouvons bien ainsi avec chacune des méthodes pour les 4 prédicteurs les valeurs suivantes :

D2	0,369

D7	0,191
D9	0,152
D14	0,455

Tableau 5.2.2.1bnlarn. fonction `predict` .*UK Data*

Cette approche ne donne donc aucune importance aux prédicteurs qui ne sont pas parents de la variable d'intérêt dans le graphe dirigé obtenu.

Voici les résultats obtenus avec le package `pcalg` et la méthode `IDA`.

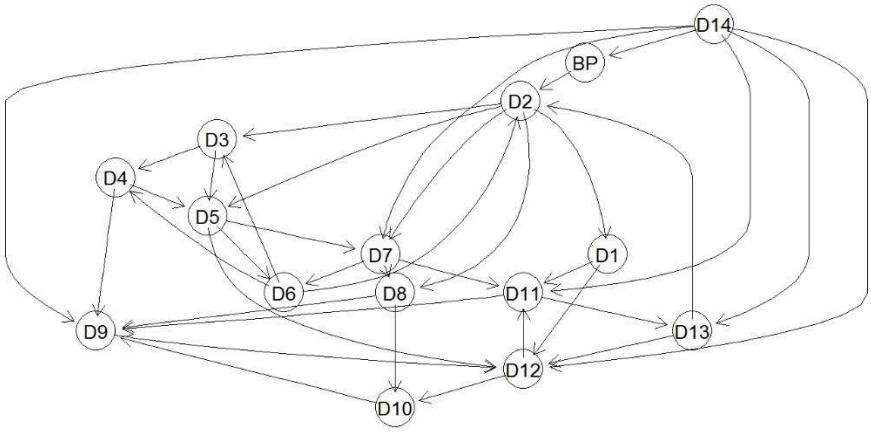


Figure 5.5.2.15 Graphe `pcalg`. UK Data

L'application de la méthode `pcalg-IDA` donne les résultats suivants :

	Pcalg/IDA UK Data
D1	0,1983995
D2	0
D3	0,2163816
D4	0,1423362
D5	0,158301

D6	0,1731028
D7	0,1809691
D8	0,1141102
D9	0,1498155
D10	0,02148003
D11	0,2026093
D12	-0,0093911
D13	0,1840371
D14	0,9427537

Tableau 5.2.2.2. `pcalg`. *UK Data.*

Dans ce cas la variable D2, qui est dans le graphe généré un enfant de la variable d'intérêt BP ne se voit allouer aucune importance. L'utilisation d'IBN livre un graphe moyen comprenant de nombreux arcs bidirectionnels, avec un nombre d'arcs plus élevés que le graphe le plus parcimonieux ci-dessus.

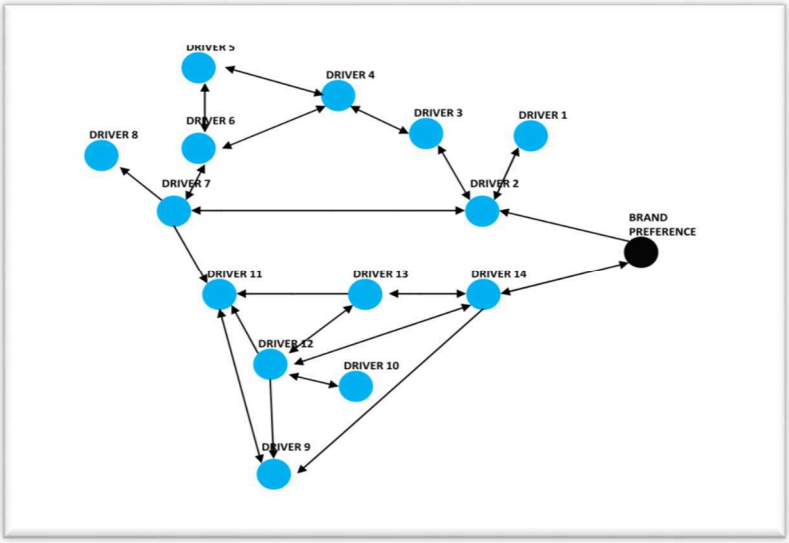


Figure 5.5.2.15. Données UK Data. Graphe avec IBN.

Là encore pour rendre le graphe plus interprétable il est utile d'avoir comme avec IBN des critères de sélection des arcs (comme the seuil d'apparition des arcs dans les tirages bootstrap pour IBN) ou d'utiliser des white et black lists (bnlearn) ou encore de fixer des seuils sur la force des arcs. Les importances calculées via IBN sont ordonnées ci-après :

	IBN
Driver 14	0,66
Driver 2	0,29
Driver 7	0,21

Driver 13	0,21
Driver 3	0,15
Driver 1	0,12
Driver 11	0,12
Driver 12	0,11
Driver 4	0,10
Driver 6	0,10
Driver 9	0,10
Driver 8	0,09
Driver 5	0,08
Driver 10	0,05

Tableau 5.5 2.3. Importances Calculées. UK Data. Simulations avec IBN.

Les variables 14, 2, 7 et 13 sont classées par cette méthode comme importantes ce qui est cohérent avec les résultats de lmg-Shapley et avec les résultats de VSURF comme présentés précédemment. La proximité avec les valeurs obtenues avec *pratt* et les B_s de l'OLS est illustrée ci-après :

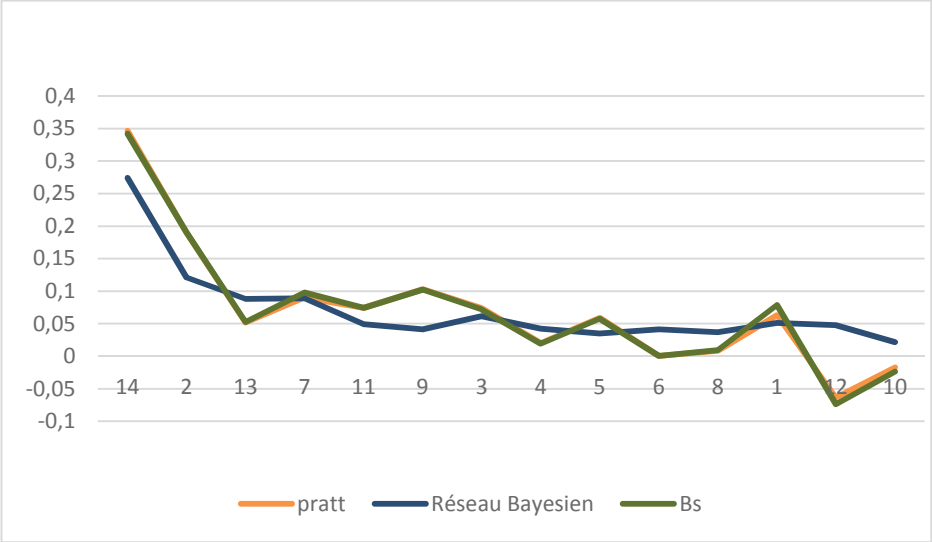


Figure 5.5.2.16. Comparaison *pratt*-réseau bayésien-OLS. UK Data. Importances normalisées.

Les différentes méthodes allouent toutes à la variable D14 la plus grande valeur d'importance, mais les résultats apparaissent effectvement différents tant dans l'ordre des variables que dans les valeurs selon le type de méthode utilisée.

Données *swiss* 182 (non quadratisées)

Ces données ont déjà été présentées.

Les quatre méthodes par contraintes de bnlearn rencontrent la même situation où la v-structure *Fertility*>>>*Education*<<<*Examination* n'est pas mise en œuvre car l'arc entre Education et Fertility apparaît dans les deux sens en raison d'une v-structure adjacente *Agriculture*>>>*Fertility*<<<*Education*.

Ces quatre algorithmes délivrent exactement le même graphe avec huit arcs.

Les autres algorithmes délivrent entre 7 et 18 arcs, le plus parcimonieux étant obtenu avec Restricted Maximisation et le plus dense avec Hiton Parents and Children.

Les graphes sont présentés ci-après :

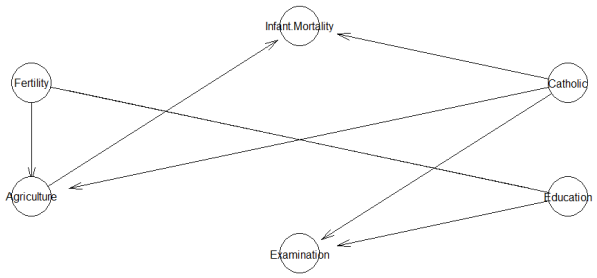


Figure 5.5.2.17. Données *swiss* 182 (non quadratisées). Graphe GS.

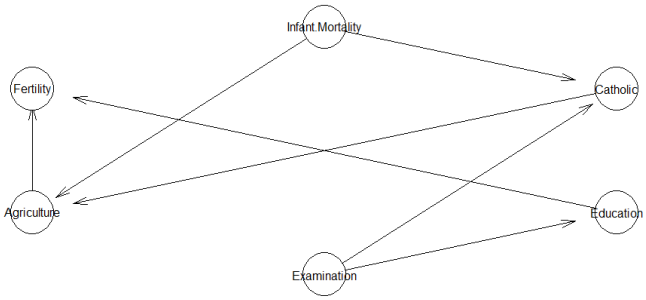


Figure 5.5.2.18. Données *swiss* 182 (non quadratisées). Graphe Restricted Maximisation.

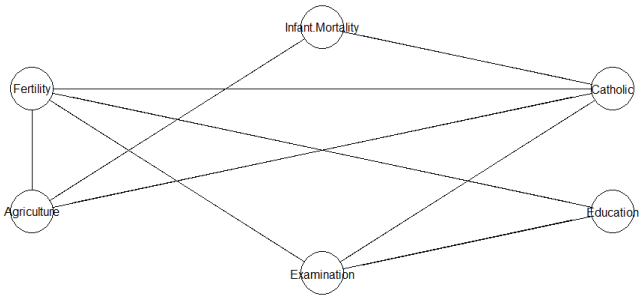


Figure 5.5.2.19. Données *swiss* 182 (non quadratisées). Graphe Hiton Parents and Children.

L'utilisation de bnlearn avec Hill Climbing fournit le graphe suivant :

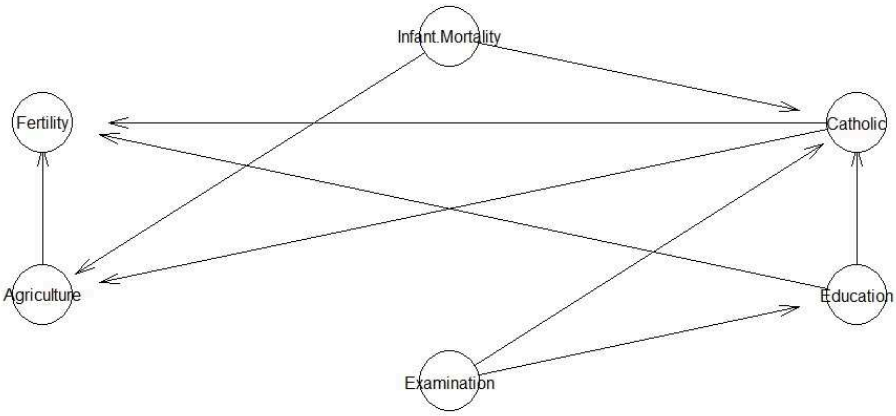


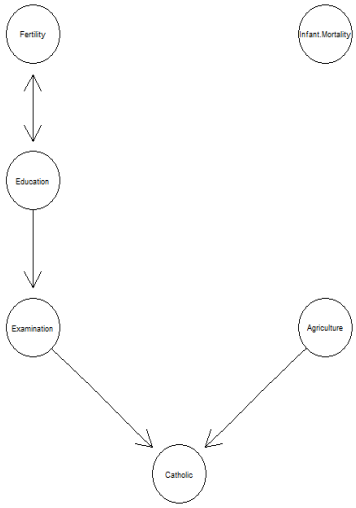
Figure 5.2.2.20. `bnlearn`. Hill Climbing. Données Swiss *182.*

Et voici les importances calculées par `prédict`, qui correspondent comme attendu aux trois coefficients de la régression de la variable d'intérêt « Fertility » sur ses trois parents dans le réseau :« Agriculture », « Education » et « Catholic ». Notons que la variable influente (et corrélée à « Education ») « Examination » ne se voit ici allouer aucune valeur d'importance car elle n'est pas parent de la variable à prédire dans le réseau obtenu.

Agriculture	0,1259
Examination	0
Education	-0,6429
Catholic	0,0516
Infant.Mortality	0

Tableau 5.2.2.4. bnlearn. Hill Climbing. *Swiss 182.*

L'utilisation de `pcalg` donne le graphe suivant :



Graphe 5.2.2.20. `pcalg`. Données *swiss* 182.

Les importances calculées avec `pcalg` sont présentées ici selon des DAG possibles :

`pcalg`	DAG 1	DAG 2
Agriculture	0,2098016	0,2098016
Examination	-0,4617253	-0,4617253
Education	-0,680789	0
Catholic	0,01199811	0,01199811
Infant Mortality	0,1843014	0,1843014

Tableau 5.2.2.5 `pcalg`. Importances IDA. Swiss 182.

Pour Agriculture, l'importance IDA est le coefficient de la régression de Fertility sur Agriculture,

Pour Examination, l'importance IDA est le coefficient d'Examination dans la régression de Fertility sur Examination et Education.

Pour Education l'importance IDA est le coefficient d'Education dans la régression de Fertility sur Education (DAG 1) ou bien est égal à zéro (DAG 2).

Pour Catholic, l'importance IDA est le coefficient de Catholic dans la régression incluant aussi les deux parents Examination et Agriculture

Pour Infant Mortality, l'importance IDA est le coefficient de la régression de Fertility sur Infant Mortality.

Voici maintenant les résultats obtenus avec IBN :

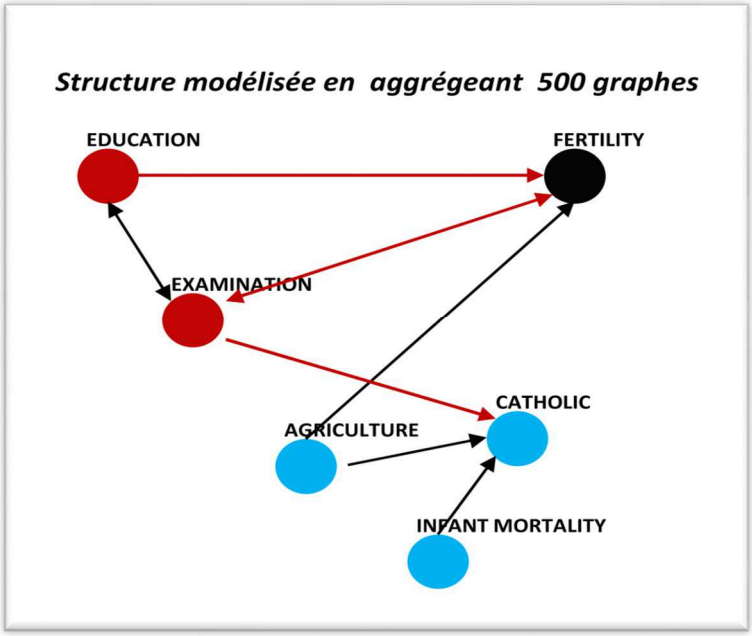


Figure 5.5.2.21. Données *swiss* 182 (non quadratisées). Graphe BBN.

Les importances calculées par la simulation via IBN sont présentées ci-après :

	IBN
Agriculture	0,41
Examination	-0,43
Education	-0,72
Catholic	0
Infant Mortality	0

Tableau 5.2.2.6. Importances Calculées. Swiss 182 (non quadratisées). Simulations avec IBN.

La simulation avec IBN restitue bien la corrélation négative avec la variable à prédire pour Examination et Education. Le réseau restreint construit avec IBN ne retient aucun arc induisant une influence des variables Catholic et Infant Mortality sur Fertility et la simulation par conséquent conduit à ne pas quantifier d'importance pour Catholic et Mortality qui sont donc *de facto* exclues du modèle.

La comparaison entre les b's de l'OLS et les coefficients générés par les réseaux est illustrée ci-dessous.

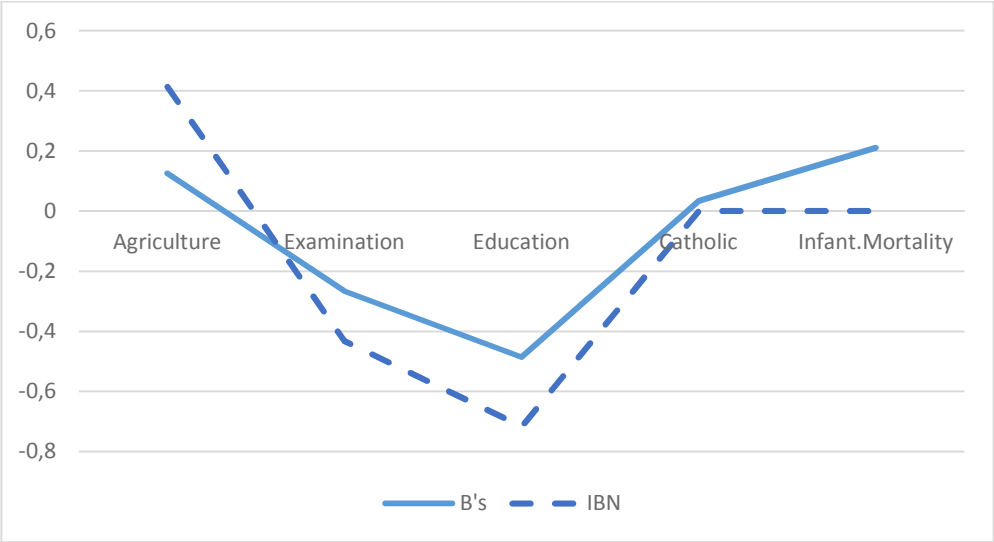


Figure 5.5.2.22. Importances IBN- b's de OLS. Données *swiss* 182 (non quadratisées).

Données *credit*

Les quatre méthodes par contraintes génèrent des graphes allant de 23 à 31 arcs, les deux méthodes de score génèrent des graphes avec 27 arcs. Le graphe le plus dense est obtenu avec Hiton Parents and Children pour 88 arcs, et les méthodes Chow-Liu et ARACNE donnent deux graphes identiques (ceci peut être facilement testé avec la fonction `all.equal (resX,resY) de bnlearn`) de 18 arcs.

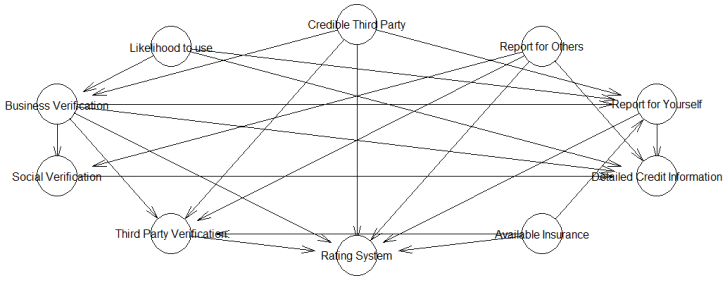


Figure 5.5.2.23. Données Credit. Graph GS.

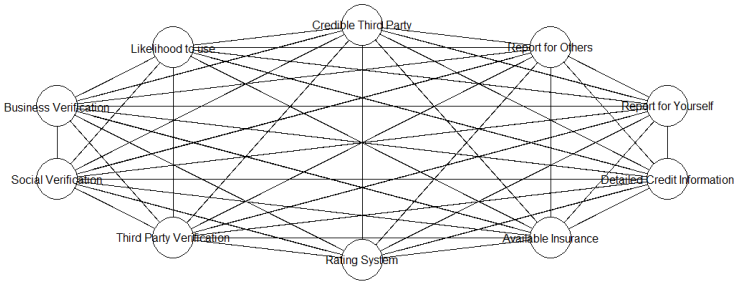


Figure 5.5.2.24. Données Credit. Graph Hiton Parents and Children.

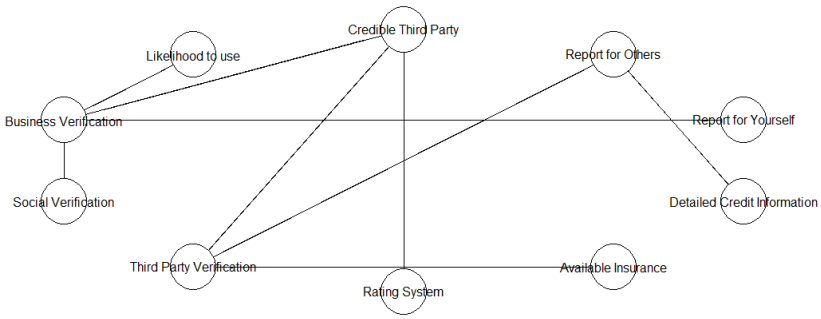


Figure 5.5.2.25. Données Credit. Graph ARACNE.

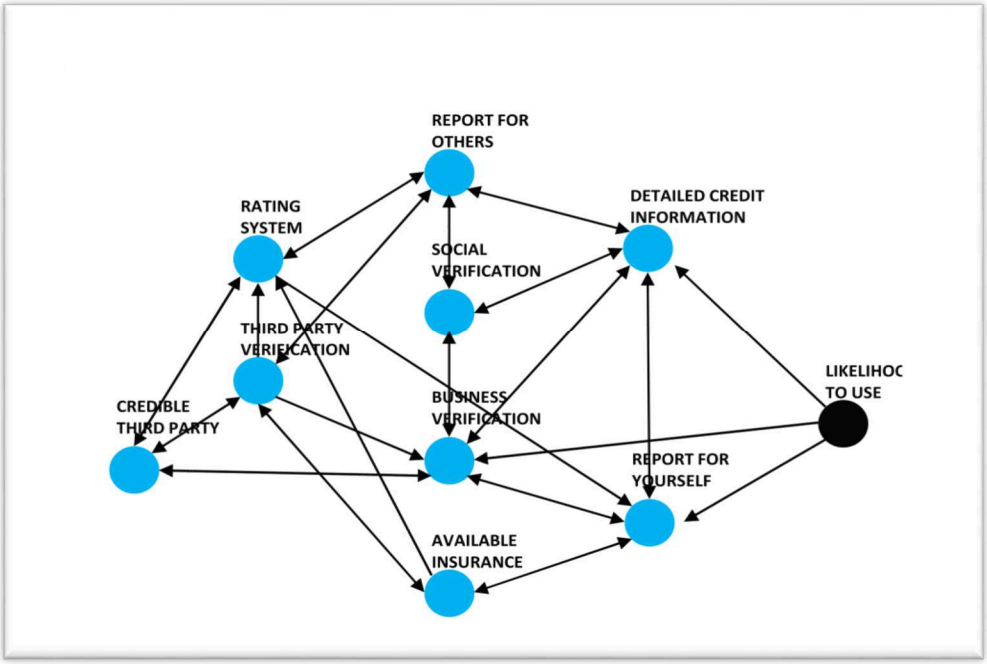


Figure 5.2.2.26. Réseau bayésien. Données Crédit. Graphe IBN.

	Réseau Bayésien
Business Verification (Driver 1)	0,21
Social Verification (Driver 2)	0,14
Third Party Verification (Driver 3)	0,08
Rating System (Driver 4)	0,05
Available Insurance (Driver 5)	0,05
Detailed Credit Information (Driver 6)	0,26
Report for Yourself (Driver 7)	0,15
Report for Others (Driver 8)	0,08
Credible Third Party (Driver 9)	0,13

Tableau 5.2.2.7. Importances calculées. IBN. Données Credit.

Les 3 premiers leviers identifiés par les réseaux bayésiens sont les leviers 6, 1 et 7 qui sont les trois parents de la variable à prédire. (variables « Detailed Credit Information », « Business Verification », « Report for yourself »). A noter également que les coefficients générés par les réseaux bayésiens dans ce cas sont tous positifs ce qui n'est pas le cas dans la régression linéaire pour Rating System et Third Party Verification (p value <0,01 pour Rating System).

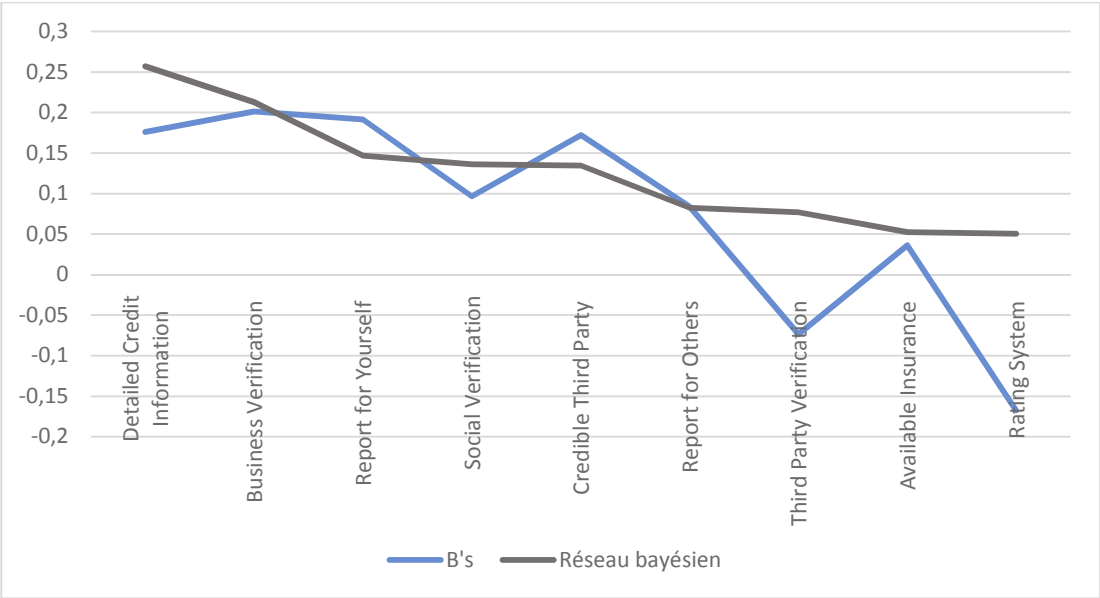


Figure 5.5.2.25. Comparaison b's de l OLS-Réseau bayésien. Données Crédit.

5.3 Commentaires sur l'utilisation des réseaux bayésiens

L'apport des réseaux bayésiens tels qu'utilisés ici, c'est-à-dire sans les variantes possibles apportés par les options d'utilisation en termes d'optimisation des graphes et de restrictions ou imposition de certains arcs est :

- de fournir une importance positive ou négative (exemple négatif avec les données *swiss*182)
- de proposer une structure de référence stable (grâce au bootstrap qui peut aussi être mis en œuvre avec bnlearn comme avec IBN) qui peut permettre de fonder un choix de modèle causal
- de permettre une modélisation des impacts de chaque prédicteur en tenant compte des relations entre variables formalisées par les graphes générés et qui permet en fait de ne pas rencontrer les inconvénients entraînés par la colinéarité dans la régression multiple dans la mesure ou des variables corrélées seront placées sur des arcs différents ou successivement sur un même arc mais que dans la simulation elles auront tendance dans de nombreux cas à être créditées d'importances proches, évitant ainsi le phénomène de forte différence entre coefficient de régression intervenant sur des variables très corrélées dans le cas de la régression multiple.

La proximité entre les coefficients de la regression simple et les importances calculées ci-dessus avec le réseau bayésien construit par indépendances conditionnelles peut être interprétée comme la conséquence du fait que ces deux approches consistent à utiliser un modèle direct et simple dans le cas de la régression linéaire ou structuré dans le cas d'un réseau bayésien mais qu'au total ces deux approches sont assez voisines comparées aux mesures d'importance reliées à des valeurs de variance (comme lmg-Shapley , PMVD , weifila) ou assimilables à des variances comme dans le cas de l'importance MSE des forêts aléatoires.

La comparaison avec l'importance Pratt a aussi été faite dans les cas particuliers présentés car les termes même de la décomposition de Pratt peuvent être associés pour chacun des termes à un réseau simple où le prédicteur j est le parent de la variable à prédire et de tous les autres prédicteurs, et où chacun des autres prédicteurs est aussi le parent de la variable à prédire.

Le réseau bayésien effectue donc intrinsèquement une sélection de relations conditionnelles et une quantification des effets indirects. En fait cette méthode est plus proche des méthodes fondées sur modélisations structurelles comme par exemple les méthodes SEM (Structural Equations Models) que sur les attributions d'importances étudiées aux chapitres 2 et 3.

Les réseaux bayésiens peuvent être utilisés comme outils de simulation ou d'aide à la décision en vue de postuler une structure de causalité. Il existe une grande variété de calculs de l'importance relative des prédicteurs, et il ne ressort pas de mesure de référence unanime de l'importance.

Les diagrammes IPA produits utilisent une mesure d'importance qui est assimilable à une corrélation bi-varie entre les facteurs sélectionnés et la variable à prédire. Cette approche permet donc de choisir des thèmes de priorités et ensuite de tenter d'identifier les leviers d'actions pratiques susceptibles d'être les plus efficaces soit en travaillant à partir des données soit en prenant en considération la capacité d'amélioration par la force du lien quantifié avec a variable à prédire soit si des variables observées s'avèrent basses en performance et présentent donc un potentiel d'amélioration.

Dans beaucoup d'applications l'impact d'un prédicteur sera calculé avec un réseau bayésien grâce à une propagation via les arcs des changements simulés des valeurs du prédicteur. Ceci offre de nombreuses variantes de calculs qui ajoutent d'autant à la diversité des réseaux plausibles.

Une autre façon de présenter l'importance peut être aussi de constituer une « tour de Hanoï » en considérant les amplitudes de variations des valeurs de la variable à prédire conditionnellement aux amplitudes constatées ou estimées réalistes des valeurs des prédicteurs.

Relevons que les calculs faits en prenant seulement les observations d'un prédicteur à un certain niveau (par exemple niveau maximum de satisfaction sur un prédicteur particulier) pour calculer l'impact via le réseau découvert sur une variable à prédire conduisent à n'utiliser qu'une partie des observations. L'approche consistant à simuler une variation de la distribution et à propager l'évolution des distributions conditionnelles dans le réseau présente l'avantage d'utiliser l'ensemble des observations et semble plus convaincante dans le cas des études de marchés qui sont parfois exploitées avec un nombre relativement restreint d'observations (200).

Ticehurst et al. (2010) ont étudié l'application des réseaux bayésiens pour estimer l'importance de prédicteurs dans des applications liées à l'agriculture mais leurs résultats et recommandations sont utiles ici. Tout comme dans les exemples précédents les prédicteurs très influents ressortent de façon cohérente entre les méthodes (Réseau, OLS). Mais Ticehurst et al. soulignent que la situation est plus variable pour les prédicteurs de moindre importance. Ils

relèvent aussi comme identifié plus haut la variabilité des résultats selon différentes options de calcul du réseau bayésien et mettent en avant le bénéfice d'une approche combinant les méthodes statistiques classiques et les réseaux bayésiens, permettant à la fois la prise en compte des jugements d'expert, la découverte de relations conditionnelles tout en intégrant les analyses obtenues par les moyens statistiques pour orienter la recherche du réseau.

De la même façon Zheng et al. (2009) ont étudié une approche qui intègre les réseaux bayésiens et la modélisation par équation structurelle, en identifiant des variables latentes puis en structurant analysant à la fois le modèle externe (ou de mesure) et le modèle interne (ou structurel) grâce aux réseaux bayésiens.

De ce point de vue, les applications dans le domaine des études de marché sont en termes de taille d'échantillon et de nombre de variables fort différentes d'applications qui ont engendré un grand succès des réseaux bayésiens comme par exemple le scoring ou le diagnostic.

Une perspective pourrait donc être d'utiliser plus formellement dans les études de marchés les réseaux bayésiens comme un outil d'aide à la décision. Ainsi les graphes obtenus, qui peuvent comme le montrent les outils présentés ci-dessus être des DAG mais aussi des graphes partiellement orientés voire non orientés, peuvent servir à élaborer un modèle structurel puis avec le recours à des équations de modélisation permettent de procéder ensuite à une modélisation statistique classique le cas échant en incorporant un jugement d'expert.

Enfin en tout état de cause l'inférence causale doit être présentée avec de grandes précautions dans la communication avec les clients et la grande variété des graphes possibles devrait conduire à une prudence dans la justification de la construction d'un modèle à partir des seules données. En outre la construction des graphes est souvent accompagnée d'hypothèses telles que la direction des arcs, l'imposition ou au contraire l'interdiction de certains arcs etc.

Enfin l'utilisation des réseaux bayésiens devrait être accompagnée des réserves d'usage quand la taille des échantillons devient retreinte, sachant qu'il arrive dans les applications des études de marchés de réaliser de tels modèles avec des échantillons de l'ordre de seulement 250 observations.

Chapitre 6. Conclusions et perspectives

6.1 Un sujet de recherches actif

L'analyse par la régression linéaire est une des méthodes statistiques les plus utilisées dans de nombreux domaines d'applications. Dans le cas des études de marchés et d'opinion, elles sont d'ailleurs la première méthode recommandée par exemple dans des manuels techniques de référence. Cependant la corrélation entre les réponses obtenues lors des enquêtes combinée avec la taille parfois limitée des échantillons conduit à un effet de multi-colinéarité se traduisant par des coefficients parfois négatifs et contraires à l'intuition des praticiens, ainsi qu'à des intervalles de confiance élevés et une instabilité des valeurs des coefficients et donc à une non reproductibilité des résultats. C'est pourquoi de nombreuses méthodes alternatives ont été étudiées et mises en pratique.
Le sujet reste à ce jour actif en termes de recherches et de publications, et aucune de ces méthodes n'a aujourd'hui fait l'unanimité. Dans un article de synthèse, Grömping (2015) souligne ainsi que bien qu'il existe de nombreuses méthodes de quantification de l'importance des prédicteurs, il n'y aurait pas encore de justification théorique convaincante pour les légitimer et qu'elles comportent toutes un aspect heuristique. Grömping indique cependant que certaines de ces méthodes restent utiles dans l'évaluation de l'importance relative, en l'absence d'une théorie experte du sujet modélisé, comme d'un modèle structurel postulé *a priori*. Elle appelle de ses vœux le développement de solutions logicielles plus conviviales que les packages R disponibles en vue de mettre des méthodes plus sophistiquées que la t-statistique des coefficients ou les simples coefficients standardisés de la régression à la disposition d'un plus grand nombre d'utilisateurs dépassant le cercle des experts, mais craint que l'absence de consensus scientifique sur les méthodes ne freine le développement d'une solution largement commercialisée.

En ce qui concerne l'estimation de l'importance à partir de la variance expliquée par le modèle linéaire nous avons identifié dans les publications :

- trois méthodes d'allocation (c'est-à-dire que la somme des allocations n'est pas en général égale au R^2) : les méthodes *first, last* et *betasq*.
- et neuf méthodes de décomposition :
    - la méthode *pratt*,
    - Trois méthodes issues de la théorie des jeux (*lmg/Shapley*, *pmvd* et *Owen value)*
    - cinq méthodes utilisant des bases orthogonales de l'espace des prédicteurs (*CAR scores,* décomposition de Green et al, et décomposition de Genizi-Johnson, décomposition de Fabbris et

aussi décomposition orthogonale via les composantes principales (DCP) mentionnée par Fabbris (1980)).

Ces méthodes ont à l'exception de celle de *green*, *fabbris* et DCP ont fait l'objet de packages déposés dans R entre 2006 et 2013. L'utilisation des réseaux bayésiens et des forêts aléatoires pour quantifier l'importance des prédicteurs a aussi été étudiée avec plusieurs solutions logicielles.

6.2 Principaux résultats

Les principaux résultats obtenus dans cette recherche sont résumés ci-après.

Une démonstration de l'égalité entre la décomposition de Johnson et lmg-Shapley dans le cas de deux prédicteurs a été établie en utilisant une approche trigonométrique ainsi que l'identification d'un majorant de la somme des firsts plus de la somme des last dans le cas d'un nombre quelconque *p* de prédicteurs. La proximité ainsi calculée entre *lmg-Shapley* et la méthode de Johnson nous a amené à contester l'argument mis en avant par Johnson selon lequel cette proximité serait en soi une justification de validité.

Une nouvelle méthode de décomposition de la variance appelée *weifila* pour « *weighted first last* » est proposée et a fait l'objet d'une publication en 2015. Elle conduit exactement aux mêmes résultats que *lmg-Shapley* et *johnson* dans le cas de deux prédicteurs et également à des résultats très proches de ces deux méthodes dans le cas général avec plus de deux prédicteurs. En revanche cette méthode est nettement plus simple car elle consiste à allouer à chaque prédicteur une importance pondérée entre le *first* et le *last* (d'où le nom *weifila*). Le script R est joint en annexe et les temps de calculs sont beaucoup plus courts qu'avec *lmg-Shapley.*

Il a été aussi montré qu'avec les jeux de données typiques des applications des études de marchés, les temps de calculs deviennent prohibitifs avec la méthode *pmvd*. Il a également été constaté que le temps de calcul avec la méthode *lmg-Shapley*, qui double avec chaque prédicteur ajouté rend cette dernière plus difficile d'utilisation à partir d'une dizaine de prédicteurs.

En ce qui concerne la méthode *CAR scores* nous avons relevé trois désaccords avec une publication scientifique sur le sujet (Strimmer et Zuber, 2011) analyse qui a été communiquée aux auteurs. Le principal point porte sur le fait que les carrés des *CAR scores* des prédicteurs ne tendent pas forcément à être égaux quand la corrélation entre ces prédicteurs tend vers 1.

Il a été aussi établi que la décomposition de la variance proposée par Fabbris (1980) n'est en réalité pas identique à celles proposées par Genizi (1993) et Johnson (2000). L'assimilation de ces trois méthodes comme identiques est un point de désaccord avec Grömping (2015). Ce résultat a été établi à la fois au plan théorique avec des expressions explicites dans le cas de deux prédicteurs et a été confirmé numériquement avec plusieurs jeux de données dans le cas de plus de deux prédicteurs. Un script R permettant de calculer la décomposition de Fabbris a été écrit et est joint en annexe.

En ce qui concerne la décomposition de Green, un script R a aussi été écrit dans la mesure où ni cette décomposition ni celle de Fabbris ne figuraient dans les packages existants. Ceci a mis en évidence une différence dans les résultats calculés avec ceux publiés par Grömping (2015) et conduit à contester la conclusion (Grömping 2015) selon laquelle la méthode de Green serait dans ses résultats assez éloignée de *genizi-johnson.*

En ce qui concerne les forêts aléatoires les résultats conduisent à considérer que les mérites de ces méthodes pour prendre en compte les non-linéarités et interactions sont sous-estimés. Nous avons ainsi observé que le choix de mtry=1 dans Random Forest RF-CART permet de calculer des importances relatives proches des résultats donnés par lmg-Shapley et ce sur un nombre varié de jeux de données. En reprenant les données utilisées par Grömping (2009 et 2015) sur des données quadratisées ou non (données *swiss* 182) il est apparu que RF-CART avec mtry=1 génère des importances relatives cohérentes avec *lmg-Shapley* après quadratisation des variables, mais que RF est aussi cohérent avec ces deux calculs d'importance (c'est-à-dire données quadratisées, RF et *lmg-Shapley)* même si les variables sont non quadratisées ce qui n'est pas le cas avec la méthode lmg-*Shapley*. Ceci confirme l'intérêt de l'utilisation des forêts alétoires pour la prise en compte des non-linéarités et interactions.

En contradiction avec Zuber (2011) et Grömping (2015), nous proposons en conclusion de ne recommander ni l'usage des *CAR scores*, ni celui de *lmg-Shapley* ou de *pmvd*. La recherche conduite ici conduit à recommander *weifila* plutôt que *lmg-Shapley* en raison de la simplicité et rapidité de *weifila*, pour des résultats très similaires. Si de surcroît une capacité de prise en compte des non linéarités est recherchée il est alors recommandé d'utiliser Random Forest avec mtry=1 plutôt que *lmg-Shapley*.

Mais Random Forest seul ne permet pas de définir et calculer des seuils de sélection ni pour l'interprétation ni pour la prévision. Dans une perspective de sélection de variables le package VSURF apporte une solution et a été mis en œuvre dans cette recherche, conduisant là encore à recommander plutôt que *lmg-Shapley* l'utilisation de VSURF avec mtry=1 pour l'étape de sélection de variables et mtry=défaut pour l'interprétation. Et plutôt que *pmvd* ou *pratt* comme le propose Grömping (2015) en alternative au modèle linéaire, nous recommandons ici VSURF avec mtry=défaut pour l'étape de sélection de variables et mtry=défaut pour la prédiction.

Les méthodes *lmg-Shapley* et *pmvd* n'ayant ni le mérite de la simplicité et de la rapidité ni la capacité de traiter les non linéarités et interactions apparaissent donc finalement de moindre intérêt, et il paraît plus pertinent d'utiliser directement *weifila* ou d'avoir recours aux techniques des forêts aléatoires qui mériteraient un usage plus large dans les applications d'études de marché, sans pour autant rejeter l'intérêt des approches structurelles et de l'apport de modèles conceptuels.

6.3 Perspectives

Ceci conduit à proposer plusieurs perspectives.

En premier lieu, les résultats obtenus sur les méthodes de décomposition de la variance (*CAR scores* et f*abbris*) et l'intérêt des forêts aléatoires pour l'estimation d'importance des prédicteurs feront l'objet d'une proposition de publication.

En termes de recherches complémentaires, une analyse comparée et détaillée de la robustesse de ces méthodes pourrait être intéressante notamment pour quantifier les variabilités respectives des résultats de ces méthodes.

En prolongement des remarques formulées par Grömping (2015) sur l'absence de solutions logicielles permettant aux « *less statistically inclined users* », pour reprendre ses termes, de bénéficier de solutions qui sont aujourd'hui seulement accessibles à des praticiens spécialisés, il pourrait être intéressant de développer des outils adaptés à une plus large diffusion assortis d'une documentation et d'un support adéquats. En particulier il serait certainement utile de proposer une solution privilégiant des calculs plus rapides que *lmg-Shapley* ou *pmvd*, et favorisant également un recours plus aisé aux forêts aléatoires.

Enfin l'analyse de l'importance des prédicteurs pourrait être explorée dans le cadre d'autres formes de réponses à des enquêtes comme les réponses binaires ou multinomiales.

Annexes

Annexe 1. Jeux de Données

- **Données *swiss***

Le jeu *swiss* est disponible directement via la console R

```
str( swiss)

'data.frame':   47 obs. of  6 variables:

 $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 .

 $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...

 $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...

 $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...

 $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...

 $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

- **Données *swiss 182***

Le jeu de données « *swiss* 182 » peut être déchargé sur le site de Princeton (http://opr.princeton.edu/archive/FileList.aspx?StudyID=10) : fichier : efswitz.dat1888

Les données sont tranformées via un script disponible auprès de Grömping (*R script to define variables : swiss_largedata_read.r)*

Les transformations sont les suivantes :

```
 swiss182 <- read.table(paste("D:/ …./efswitz.dat1888",sep="/"))

 swiss182 <-  swiss182[,c(2,20,31,33,6,19)]

colnames( swiss182) <- colnames( swiss)

 swiss182$Fertility <-  swiss182$Fertility / 10

 swiss182$Agriculture <-  swiss182$Agriculture / 10

 swiss182$Examination <-  swiss182$Examination / 1000

 swiss182$Education <-  swiss182$Education / 1000

 swiss182$Catholic <-  swiss182$Catholic / 1000

 swiss182$Infant.Mortality <-  swiss182$Infant.Mortality / 10
```

- ***Données UK Data***

Les données « UK Data sont accessibles via le lien suivant :

https://www.dropbox.com/sh/gez9fqbma9q8e0o/AACWbhulakOELt1jKY4jP8s_a?dl=0

Elles comportent 1599 observations, et 17 variables.

- **Données *auto***

Les données auto.dta sont accessibles sur le site de la société Stata (www.stata.com) url : http://www.stata-press.com/data/r10/u.html.Elles ont été utilisées ici avec R en convertissant le fichier de données grace à la fonction « foreign » de R.

```
    R Code
    library(foreign)
Auto <- read.dta("D:/"parth to your file"/auto.dta")
```

Ces données comportent 74 observations et 12 variables. Elles ont été utilisées avec comme variable à prédire la variable mpg (c3).

- **Données *credit***

Les données *credit* sont issues d'un exemple réel.

Elles sont accessibles via le lien suivant :

https://www.dropbox.com/sh/gez9fqbma9q8e0o/AACWbhulakOELt1jKY4jP8s_a?dl=0

Le nom du fichier de données est study 24238.csv et le script R permettant d'effectuer tous les calsuls d'importance est aussi inclus (ScriptCompletCreditR). Il est nécessaire d'installer et charger `relaimpo` (non US version), `svd` et `randomForest`.

Le script permet de charger le nom des variables.

Les données comportent 499 observations , la variable à prédire est « Likelihood to Use » et les 9 prédicteurs sont :

Business Verification , Social Verification ,Third Party Verification , Rating System, Available Insurance , Detiles Credit Information ,Report for yourself , Report for Others , Credible Third Party.

Annexe 2. Scripts R utilisés

- **weifila** (sans utiliser relaimpo) (Example UK Data)

```
# UKData:leresultatderead.table
UKData<-read.table("UKData.csv",sep=",",header=TRUE)
# str pour 'structure'
y <- as.matrix(UKData[,3])
x <- as.matrix(UKData[,c(4:17)])
yx <- as.data.frame(UKData[,c(3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)])
# function Weifila
weifila <- function(formula, data=NULL) {
  full.model <- lm(formula, data)
  last <- (drop1(full.model)["Sum of Sq"] / sum(anova(full.model)["Sum Sq"]))[-1,1]
  n <- length(last)
  first.frame <- model.frame(formula, data)
  first <- cor(first.frame)[1,-1]^2
  r.squared <- summary(full.model)$r.squared
  L <- sum(last)
  F <- sum(first)
  if ((L <= r.squared) && (r.squared <= F)) {
    W <- array(dim=n)
    for (i in 1:n) {
      W[i] <- last[i]*(F-r.squared)/(F-L) + first[i]*(r.squared-L)/(F-L)
    }
  } else if ((F <= r.squared) && (r.squared <= L)) {
    W <- array(dim=n)
    for (i in 1:n) {
      W[i] <- last[i]*(r.squared-F)/(L-F) + first[i]*(L-r.squared)/(L-F)
    }
  }
```

```
  if (exists("W")) {
    return(list(L=L, r.squared=r.squared, F=F, W=W))
  }
  else { cat("Special Case\n")
    return(list(L=L, r.squared=r.squared, F=F))
  }
}
# call weifila : le résultat va dans la liste 'output', on utilise le raccourci de formule ~.
# BP est la variable Y dans le dataframe yx
prt<- proc.time()
output <- weifila(BP~.,data=yx)

# call weifila : on extrait les poids (élément W)
Weights <- weifila(BP~.,data=yx)$W
print(output)
```

- **Calculs d'importance**

  Ce script regroupe l'ensemble des décompositions de la variance en utilisant relaimpo, randomForest et svd.

```
# relaimpo
crf <- calc.relimp(y~x, data=yx, type = c("lmg", "pmvd", "last", "first", "betasq",
"pratt","genizi","car"), rela = FALSE )
relimp <- cbind(crf@lmg, crf@pmvd, crf@last, crf@first, crf@betasq,
crf@pratt,crf@genizi,crf@car)
colnames(relimp) <- c("lmg", "pmvd", "last", "first", "betasq", "pratt","genizi","car")
# RF regression
incMSE <- matrix(c(0),nrow=ncol(x),ncol=ncol(x))
for (i in 1:ncol(x)){
       rf <- randomForest(x, y, ntree=250, importance=TRUE, mtry=i)
       incMSE[,i] <- rf$importance[,1]}
colnames(incMSE) <-paste(rep("RF : mtry",ncol(incMSE)),c(1:ncol(incMSE)),sep="")
all <- cbind(relimp, incMSE)
allnorm<-all%*%diag(1/colSums(all))
```

```
colnames(allnorm)<-colnames(all)

# Formule de Fabbris U de Fabbris est V dans SVD et V dans Fabbris est U dans SVD

x<-scale(x, center=TRUE,scale=TRUE)

y<-scale(y,center=TRUE,scale=TRUE)

svd<-svd(x)

# La décomposition est bien  x=UdeltaV' dans les notations du package svd , ce qui correspond
à x=Vlambda(1/2)Uprime dans les notations de l'article de Fabbris.

# calcul de gamma cf formule de Fabbris (4)

gamma<-(svd$v)%*%diag(1/svd$d)%*%(t(svd$v))

gamma2<-gamma*gamma

# calcul de poidsgreen correspondant à la formule (6) de Fabbris

sc<-colSums(gamma2)

poidsgreen<-gamma2%*%diag(1/sc)

print(poidsgreen)

# betastar correspond à la formule (8) de Fabbris

betastar<-cor(svd$u,y)

betastar2<-betastar*betastar

# Calcul des valeurs de la décomposition de Fabbris avec la formule de "Notes 2" de l'article
de Fabbris

U2<-svd$v*svd$v

Fabbris<-U2%*%betastar2

colnames(Fabbris)<-c("fabbris")

# calculs des CAR scores , carrés des corrélations Z'y , cad UV'y  notations de Fabbris ,
donc vbetastar

carF<-svd$v%*%betastar

car2F<-carF*carF

VprimeX<-t(svd$u)%*%x

VprimeX2<-VprimeX*VprimeX

ZprimeX<-svd$v%*%(t(svd$u)%*%x)

ZprimeX2<-ZprimeX*ZprimeX

# betagenizi est Z'y

 betaGenizi<-svd$v%*%betastar

betaGenizi2<-betaGenizi*betaGenizi

# Calcul corréct de green
```

```
green<-poidsgreen%*%betaGenizi2
colnames(green)<-c("green")
# Calcul de Gromping
greenGromping<-gamma2%*%betaGenizi2
colnames(greenGromping)<-c("greenG")
# betastar est V'y
betastar<-cor(svd$u,y)
betastar2<-betastar*betastar
U2<-svd$v*svd$v
# Calcul de la Décomposition de Fabbris
Fabbris<-U2%*%betastar2
colnames(Fabbris) <- c("fabbris")
# Calcul des CAR scores pour vérification
carF<-svd$v%*%betastar
car2F<-carF*carF
 colnames(car2F) <- c("CAR svd")
colnames(car2F)<-c("CAR svd")
#calcul direct de Genizi
GeniziW<- ZprimeX2%*%betaGenizi2
GeniziW<-GeniziW/(nrow(x)-1)
colnames(GeniziW) <- c("genizi svd")
# DCPW : calcul direct de la Decomposition via orthogonalisation sur composantes principales
DCPW<-VprimeX2%*%betastar2
DCPW<-DCPW/(nrow(x)-1)

colnames(DCPW) <- c("DCPrinc.")
# calculs de weifila via relaimpo
f<-sum(relimp[,4])
l<-sum(relimp[,3])
R2<-lm$r.squared
clast<-((f-R2)/(f-l))
cfirst<-((R2-l)/(f-l))
```

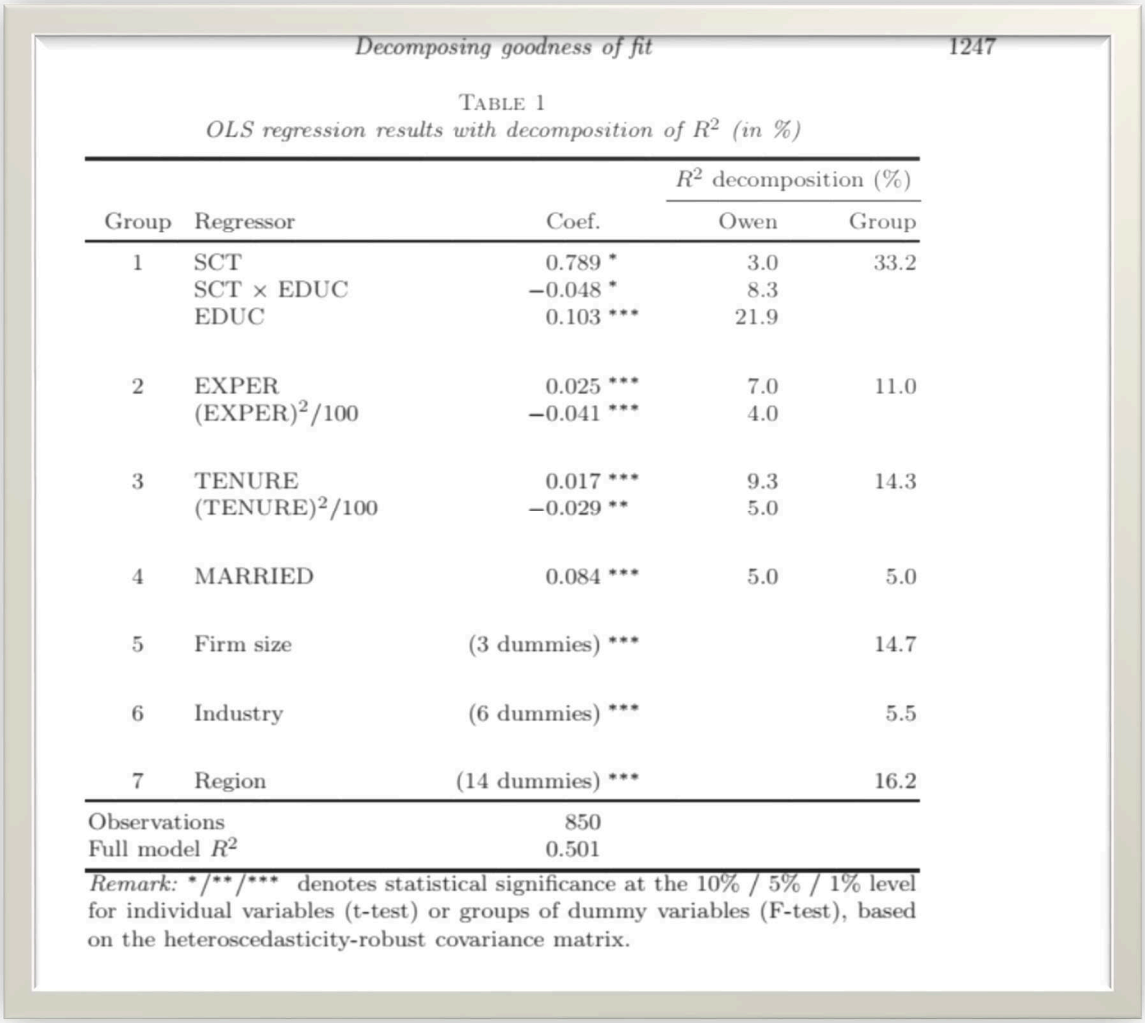
```
fila<-relimp[,3:4]
print(fila)
weifila<-clast*fila[,1]+cfirst*fila[,2]
weifila<-as.matrix(weifila)
colnames(weifila)<-c("weifila")
relimposvd<-cbind(relimp,Fabbris,green,greenGromping,car2F,GeniziW,DCPW,weifila)
print(relimposvd)
relaimposvdnorm<-relimposvd%*%diag(1/colSums(relimposvd))
print(relaimposvdnorm)
relimposvdrf<-cbind(all,Fabbris,green,greenGromping,car2F,GeniziW,DCPW,weifila)
print(relimposvdrf)
relaimposvdrfnorm<-relimposvdrf%*%diag(1/colSums(relimposvdrf))
colnames(relaimposvdrfnorm)<-colnames(relimposvdrf)
print(relaimposvdrfnorm)
```

- **Réseaux bayésiens**

```
library(bnlearn)
data(gaussian.test)
BNET<-gaussian.test
str(BNET)
res1=gs(BNET)
res2=iamb(BNET)
res3=fast.iamb(BNET)
res4=inter.iamb(BNET)
res5=hc(BNET)
res6=tabu(BNET)
res7=mmhc(BNET)
res8=rsmax2(BNET)
res9=mmpc(BNET)
res10=si.hiton.pc(BNET)
res11=chow.liu(BNET)
res12=aracne(BNET)
print(res1$arcs)
```

Annexe 3. Owen Value.

Exemple présenté par Hüttner en Saunder (2011)



Decomposing goodness of fit 1247

TABLE 1
OLS regression results with decomposition of R^2 (in %)

Group	Regressor	Coef.	R^2 decomposition (%) Owen	Group
1	SCT	0.789 *	3.0	33.2
	SCT × EDUC	−0.048 *	8.3	
	EDUC	0.103 ***	21.9	
2	EXPER	0.025 ***	7.0	11.0
	(EXPER)^2/100	−0.041 ***	4.0	
3	TENURE	0.017 ***	9.3	14.3
	(TENURE)^2/100	−0.029 **	5.0	
4	MARRIED	0.084 ***	5.0	5.0
5	Firm size	(3 dummies) ***		14.7
6	Industry	(6 dummies) ***		5.5
7	Region	(14 dummies) ***		16.2
Observations		850		
Full model R^2		0.501		

Remark: */**/*** denotes statistical significance at the 10% / 5% / 1% level for individual variables (t-test) or groups of dummy variables (F-test), based on the heteroscedasticity-robust covariance matrix.

Voici une vue plus détaillées des variables prises en compte et des résultats :

```
. xi: rego lwage sct01 sct01_educ educ (detail) \ exper exper2 (detail) \ tenur
> e tenure2 (detail) \ married \  i.firmsize \  i.industry \  i.state ,  vce(ro
> bust)
i.firmsize        _Ifirmsize_1-4      (naturally coded; _Ifirmsize_1 omitted)
i.industry        _Iindustry_1-7      (naturally coded; _Iindustry_1 omitted)
i.state           _Istate_1-15        (naturally coded; _Istate_1 omitted)
```

Gr	Regressor	Coef.		Std.Err.	P>|t|	Ind. %R2	Group %R2
1	sct01	.7488771	*	.3948335	0.058	3.0150	33.2328
	sct01_educ	-.0484895	*	.0288154	0.093	8.2994	
	educ	.1033054	***	.0166554	0.000	21.9184	
2	exper	.0253286	***	.0058137	0.000	6.9608	11.0005
	exper2	-.00041	***	.000128	0.001	4.0397	
3	tenure	.017496	***	.0044985	0.000	9.3083	14.3193
	tenure2	-.0002857	**	.0001182	0.016	5.0110	
4	married	.0838247	***	.0287177	0.004		5.0408
5	_Ifirmsize_2	.1276771	***	.0455978	0.005		14.6925
	_Ifirmsize_3	.2783466	***	.0499823	0.000		
	_Ifirmsize_4	.2911761	***	.0456677	0.000		
6	_Iindustry_2	.1384518	**	.0585293	0.018		5.4742
	_Iindustry_3	.1861953	***	.0605074	0.002		
	_Iindustry_4	.05868		.0668536	0.380		
	_Iindustry_5	.0397909		.0653681	0.543		
	_Iindustry_6	.1765137	**	.0746009	0.018		
	_Iindustry_7	.0108604		.0599852	0.856		
7	_Istate_2	.1512761		.1369602	0.270		16.2399
	_Istate_3	.0309796		.0923998	0.738		
	_Istate_4	-.0316632		.1868666	0.865		
	_Istate_5	.0517555		.0887172	0.560		
	_Istate_6	.0930638		.0924681	0.315		
	_Istate_7	.066802		.0898569	0.457		
	_Istate_8	.212286	**	.0955396	0.027		
	_Istate_9	.0713892		.091401	0.435		
	_Istate_10	-.1092712		.1078399	0.311		
	_Istate_11	-.127727		.1009002	0.206		
	_Istate_12	-.2645444	***	.092973	0.005		
	_Istate_13	-.2513599	**	.1236173	0.042		
	_Istate_14	-.3568086	***	.1255198	0.005		
	_Istate_15	-.2709405	**	.1123709	0.016		
-	Intercept	.529534	*	.2804751	0.059		
	Observations	850					
	Overall R2	0.50092					
	Root MSE	.3671009					
	F-stat. Model	24.7068	***		0.000		
	Log Likelihood	-337.988					

Huettner et Sunder (2012) ont donc utilisé l'Owen value pour des regroupements très intuitifs et logiques. Nous n'avons pas trouvé dans la pratique des études de marché d'utilisation de la décomposition d'Owen et cette méthode n'est pas non plus mentionnée dans l'article de synthèse publié par Grömping en mars 2015 (Gromping 2015)

Annexe 4. Calculs trigonométriques

Cas de deux prédicteurs

$$sr_1 = \sin(\varphi - \psi)$$
$$sr_2 = \sin(\varphi + \psi)$$
$$\beta_1 = \frac{\sin(\varphi - \psi)}{\sin(2\varphi)}$$
$$\beta_2 = \frac{\sin(\varphi + \psi)}{\sin(2\varphi)}$$
$$r_{y1} = \cos(\varphi + \psi)$$
$$r_{y2} = \cos(\varphi - \psi)$$
$$\rho_{12} = \cos(2\varphi)$$
$$\text{last}(1) = \sin^2(\varphi - \psi)$$
$$\text{last}(2) = \sin^2(\psi + \varphi)$$

$$\text{first}(1) = \cos^2(\psi + \varphi)$$
$$\text{first}(2) = \cos^2(\psi - \varphi)$$

$$\text{last}(1) + \text{last}(2) = 1 - \cos(2\psi)\cos(2\varphi)$$
$$\text{first}(1) + \text{first}(2) = 1 + \cos(2\psi)\cos(2\varphi)$$

$$\text{SV}(1) = \frac{(1 - \sin(2\varphi)\sin(2\psi))}{2}$$
$$\text{SV}(2) = \frac{(1 + \sin(2\varphi)\sin(2\psi))}{2}$$
$$\text{first}(1) = \text{SV}(1) + \frac{\cos(2\psi)\cos(2\varphi)}{2}$$
$$\text{first}(2) = \text{SV}(2) + \frac{\cos(2\psi)\cos(2\varphi)}{2}$$
$$\text{last}(1) = \text{SV}(1) - \frac{\cos(2\psi)\cos(2\varphi)}{2}$$
$$\text{last}(2) = \text{SV}(2) - \frac{\cos(2\psi)\cos(2\varphi)}{2}$$

Annexe 5. CAR scores

Code fourni par K. Strimmer. 7 janvier 2015

```
#' ---
#' title: "Difference of Squared CAR Scores"
#' output: pdf_document
#' author: Korbinian Strimmer
#' date: 7 Januar 2015
#' ---

#' # Required Software
#' To run this example you need to install the 'corpcor' R package that is
#' needed for  computing matrix power and check of positive definiteness.

#'
#' # Example regression model

#' This is an example regression model with two variables:

sbeta1 = 0.6  # standardized regression coefficients for X1
sbeta2 = 0.3  # standardized regression coefficients for X2
rho12 = 0.5   # correlation between predictor X1 and X2

#' We now compute the difference of squared CAR scores for X1 and X2
#' from this model, with rho12 varied between 0 and 1.

#' # R Function to compute the difference of the squared CAR scores for a given model

computeDifference = function(rho12=0.5, sbeta =c(0.6, 0.3) )
+ {
+   require("corpcor") # needed for computing matrix power and check
+                      # of positive definiteness
+
+   # correlation matrix among predictors  (2x2 matrix)
+   rho.xx = matrix(c(1,rho12, rho12, 1), 2, 2)
+
+   # marginal correlations: cor(Y, X1) and cor(Y,X2)
+   rho.marg = as.vector(rho.xx %*% sbeta )
+
+   # full correlation matrix (Y, X1, X2)
+   rho = rbind(c(1, rho.marg),
+               cbind(rho.marg, rho.xx))
+
+   # check if positive definite
+   if (!is.positive.definite(rho))
+   {
+     warning("Joint model is (Y, X1, X2) not positive definite!")
+   }
```

```
+
+   # CAR scores
+   carscores = as.vector( mpower(rho.xx, -1/2) %*% rho.marg )
+   # carscores = as.vector( mpower(rho.xx, 1/2) %*% sbeta ) # equivalent
+
+   # return difference of squared CAR scores
+
+   return( carscores[1]^2-carscores[2]^2 )
+ }

 #' Here is the value for the default settings:

 computeDifference()# 0.2338269
[1] 0.2338269



 #'
 #' # Difference of squared CAR scores when rho12 is varied.



 #' We vary the correlation rho12 between 0 and 1 as follows:

 rho12.vec = c(0, .1, .2, .3, .4, .5, .6, .7, .8, .9,
+               .99, .9999, .999999, .99999999, .9999999999)

 #' Compute the difference of squared CAR scores using the above function:
 results=lapply(rho12.vec,
+   function(r)
+   {
+     return( computeDifference(rho12=r) )
+   }
+ )
 results = do.call(rbind, results)
 results = cbind(rho12.vec,results)
 results
      rho12.vec
 [1,]  0.000000 2.700000e-01
 [2,]  0.100000 2.686466e-01
 [3,]  0.200000 2.645449e-01
 [4,]  0.300000 2.575636e-01
 [5,]  0.400000 2.474591e-01
 [6,]  0.500000 2.338269e-01
 [7,]  0.600000 2.160000e-01
 [8,]  0.700000 1.928186e-01
 [9,]  0.800000 1.620000e-01
[10,]  0.900000 1.176903e-01
[11,]  0.990000 3.808819e-02
[12,]  0.999900 3.818281e-03
```

[13,] 0.999999 3.818376e-04

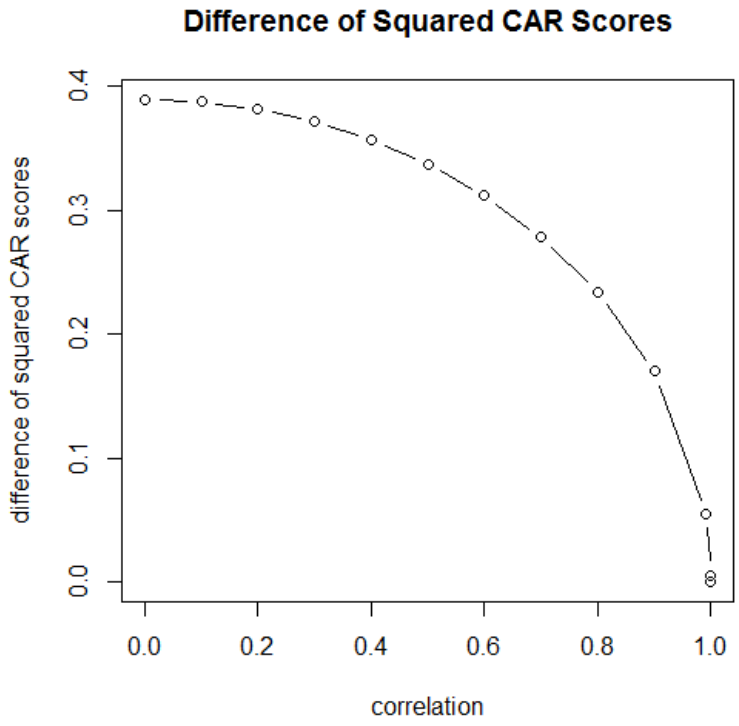
[14,] 1.000000 3.818377e-05

[15,] 1.000000 3.818378e-06

```
plot(results[,1], results[,2], type="b", main="Difference of Squared CAR Scores",
+     xlab="correlation", ylab="difference of squared CAR scores")
```

Code avec ρ=0,8 nous retrouvons bien la valeur de référence avec les coefficients correspondant à y_0 .

```
computeDifference = function(rho12=0.8, sbeta =c(0.647962743, 0.173621094) )
    rho12.vec
 [1,]  0.000000 3.897114e-01
 [2,]  0.100000 3.877580e-01
 [3,]  0.200000 3.818377e-01
 [4,]  0.300000 3.717610e-01
 [5,]  0.400000 3.571764e-01
 [6,]  0.500000 3.375000e-01
 [7,]  0.600000 3.117691e-01
 [8,]  0.700000 2.783096e-01
 [9,]  0.800000 2.338269e-01
[10,]  0.900000 1.698713e-01
[11,]  0.990000 5.497556e-02
[12,]  0.999900 5.511214e-03
[13,]  0.999999 5.511351e-04
[14,]  1.000000 5.511352e-05
[15,]  1.000000 5.511350e-06
```



Comparaison calcul R et Calcul Excel

rho12.vec	ρ	Δ CAR "R"	Δ CAR Excel
[1,]	0,000000000	0,2700000	0,2700000
[2,]	0,100000000	0,2686466	0,2686466
[3,]	0,200000000	0,2645449	0,2645449
[4,]	0,300000000	0,2575636	0,2575636
[5,]	0,400000000	0,2474591	0,2474591
[6,]	**0,500000000**	**0,2338269**	**0,2338269**
[7,]	0,600000000	0,2160000	0,2160000
[8,]	0,700000000	0,1928186	0,1928186
[9,]	0,800000000	0,1620000	0,1620000
[10,]	0,900000000	0,1176903	0,1176903
[11,]	0,990000000	0,0380882	0,0380882
[12,]	0,999900000	0,0038183	0,0038183
[13,]	0,999999000	0,0003818	0,0003818
[14,]	0,999999990	0,0000382	0,0000382
[15,]	1,000000000	0,0000038	0,0000038

Annexe 6. Article publié.

Article publié à la conférence 2015 de l'ASMDA : *The 16th Conference of the Applied Stochastic Models and Data Analysis International Society*. (Le Pirée, Grèce, 2015)

http://www.asmda.es/images/1_W-Z_ASMDA2015_Proceedings.pdf

Using Explained Variance Allocation to analyse Importance of Predictors

Henri Wallard

CEDRIC, Centre d' Etude et de Recherche en Informatique et Communications.
CNAM CEDRIC, 292 rue Saint Martin ,75141 Paris , France. Email :henriwallard@hotmail.com

Abstract. Using applications of linear regression, Market Research practitioners want to determine a ranking of predictors or a quantification of their respective importance for a desired outcome. As predictors are often correlated, regression coefficients can be difficult to use directly because they can be instable across samples and have negative values that are counterintuitive. To overcome these difficulties other methods have been proposed in the industry using squared semi partial correlation coefficients, squared zero order correlation coefficients or methods such as Shapley Value decomposition or decomposition via orthogonalisation in the space of predictors.

The proximity between the results obtained by different Variance Decomposition methods has led some authors to conclude that they are a fully valid approach. This paper will highlight theoretical reasons why these methods present similarities, offer a simple alternative new way to decompose variance but will also show the flaws and risks of relaying on Variance Decomposition for quantification of importance of predictors and why a Game Theory approach like Shapley Value can lead to misinterpretations. It will also present additional methods developed to compute β coefficients using Variance Decomposition as an intermediate step and propose recommendations for driver analysis.
Keywords: Variance Decomposition, Regression, Importance.

1 Introduction

In the field of Market Research, practitioners want to help their clients identify how to act on some factors such as quality of service or design of a product to achieve a desirable outcome such as purchase intent or satisfaction and loyalty of the customers. This is done with the desire to identify what are the best drivers of improvement, and quantify their respective impact. This is achieved through statistical modeling and simulation like in other fields such as Psychology, Social Sciences or Economics. The classic reference method used to do this Ordinary Least Square regression (OLS), but while this approach is recommended in the business literature it has some limitations. Because of sample size and design of the questions in Market research surveys, there may occur instability of coefficients. And also as questions can relate to similar topics there may be multi-collinearity and negative coefficients that are counter-intuitive. Causal assumptions and modeling options can lead to a variety of results when one wants to quantify or simulate the impact of a given action on a predictor on the desired response. In Market Research many techniques are used: Regressions, Path Models, Bayesian Belief Networks, Random Forest to name a few. This paper focusses on relative importance in the context of Linear Regression.

2 The concept of relative importance

Johnson [1] points out that the terminology of relative importance is confusing because of the many different definitions used, and introduced the term "relative weights" to define the proportion of explained variance by the linear model allocated to each individual predictor. We propose first to define relative importance in a general way.
Let us consider p random variables as predictors and y a response. We can define a Relative Importance Function as a function that $\rightarrow R^p$ associates to each predictor (defined by its index j in the set P={1,…p}) a value of importance:

$$RI : \mathrm{P} \rightarrow \square$$

In practical applications, functions of relative importance are defined using matrix calculus and polynomial functions applied to the correlation matrix, explained variance of models and bivariate or multivariate correlation coefficients. These computations are applied using the estimates of these values. As a consequence we should like Grömping [2] clearly refer to estimators of relative importance. For instance we will see later that some relative importance functions are based on a full decomposition of the explained variance. As the estimator of explained variance in Linear Regression is biased, it is impossible that all estimators of relative importance for each predictor are unbiased as their sum is actually biased. This is why in all that follows we will only discuss the properties of estimators .The topic of relative importance has been discussed in many publications since at least 1936 and a history of the use of relative importance has been presented by Johnson and Lebreton

[3]. This article will focus on relative importance evaluation based on allocation of shares of variance for linear regression. Grömping [2] gives an overview of Variance Decomposition methods. Some approaches allocate shares of Variance that can be negative and this has attracted criticism. Others propose the usage of values of relative importance that are all positive but do not add up to the total of explained variance. Lastly some methods fully decompose the explained variance across predictors. We will designate in this paper the methods that assign relative importance values to each predictor so that the sum of these relative importance adds up to the estimated R^2 as Variance Decomposition as opposed to Variance Allocation when the sum of the relative importance estimates is different from the estimated R^2.

In terms of notation we will use the following in reference to the Linear Regression Model and focus this article on allocation of shares of variance of the response y into proportions due to the p predictors X's (and errors) :

$$y = X\beta + e = \sum \beta_i X_i + e \quad (1)$$

We will work in the case of p predictors that are linearly independent, and in the case of n observations n greater than p and the n x p matrix of predictors score is of full rank p. The variance explained by the p predictors (P is the set of the p predictors) is:

$$V(P) = \sum_{i=1}^{i=p} \beta_i^2 v_i + 2 \sum_{i<j} \beta_i \beta_j \sqrt{v_i v_j} \quad (2)$$

with vi variance of Xi and ρ_{ij} coefficient of corrélation between Xi et Xj.

We can assume a regression model without intercept and w.l.o.g that all X's are centered (i.e. have expectation 0). To simplify the notations in the rest of the article we will assume, unless specified, that y and the Xj's are centered and standardized.

We have identified 8 methods published and included in R packages. A detailed documentation on 6 of these methods is available on Pr. Grömping's website [4] dedicated to resources on the relaimpo R package. Another method proposed by both Genizi [9] and Johnson [1] called here Relative Weights (RW's) and finally Zuber and Strimmer [5] CAR scores (Correlation-Adjusted (marginal) correlation) are also available via R packages (relaimpo and yhat). We will first present 3 methods of allocation and then 5 methods of decomposition and then discuss some points of difference and convergence between these approaches.

.

3 Methods for Variance Allocation

3. 1 Allocation "first".

The measures are the squared correlations of the predictors with the response:

$$first(j) = \mathrm{cov}(y, Xj)$$

When the predictors are mutually de-correlated the sum of the measures "first" adds up to the overall R^2 of the model. When this is not the case, the sum of the first (j) over all p predictors is often higher that the overall R^2 of the model.cf. Grômping [4].

3.2 Allocation "last".

This measure attributes as Relative Importance for a predictor j the increase in R^2 when predictor j is included last in the model compared to the R^2 with only the other p-1 predictors. This measure is identical to the squared semi-partial correlation $sr^2(j)$, which is sometimes presented as the amount by which the R^2 is reduced when this predictor is deleted from the regression equation. See for instance Tabachnick and Fidell [6].

3.3 Allocation "betasquared".

This relative importance measure consists in attributing as importance the square of the standardized regression coefficient. Like the measures 3.1 and 3.2 these are variance allocations as the sum of these measures for all p predictors do not in general add up to the R^2.

4 Methods of Variance Decomposition

4.1 Decomposition Hoffman-Pratt

This measure of relative importance noted pratt(j) attributes to a predictor j the product of the standardized multiple regression coefficient by the marginal correlation between the predictor j and the response When the predictors are standardized :

$$pratt(j) = \beta_j \rho_{yj}$$

From the properties of the OLS regression we can easily confirm that this measure is leads to a decomposition of the R².

$$R^2 = \text{cov}(y, \sum_j \beta_j Xj) = \sum_j \beta_j r_{yj} \tag{3}$$

4.2 Shapley Value or LMG or Average

This method has been assigned several names. See for instance Grömping [2] for an historical overview. We will call this measure here lmg(j) or SV(j). This measure is computed by averaging on all possible ordering of the p predictors the increase of the R² when the predictor j is added to the model based on the other predictors entered before j in the model. These values have been proposed by Lindeman, Merenda and Gold (1980), hence the name lmg. If we consider a game theory perspective where we assimilate the p predictors as players and define the game function of a coalition of k players as the R² achieved by the model based on these k predictors, it turns out that the application of Shapley Value to the game described above generates exactly the same values as lmg,, hence the possible notation SV(j).

Let us present below one notation and one of the ways to compute lmg. Let r be a permutation of P, this constitutes an ordering of the predictors. Each permutation r enables to define an order of entry of the predictors in the model.

Let Sj(r) be the set of predictors entered before j in the permutation r. We can compute:

$R^2(S_j(r))$ as the R² of the model including the predictors in Sj(r)

$R^2_{+j}(S_j(r))$ as the R² of the model including the predictors in Sj(r) U{j}

And define $\Delta_j(r) = R^2_{+j}(S_j(r)) - R^2(S_j(r))$

$\Delta_j(r)$ is the increase in R² when the predictor j is added to the predictors entered before j in the model with order resulting from the permutation r.

$$lmg(j) = \frac{1}{p!} \sum_r \Delta_j(r) \tag{4}$$

averaged on all 2^p permutations of the p predictors. This formula can be rewritten in different forms, combining the permutations that have the same sets Sj(r).

4.3 PMVD (Proportional Marginal Variance Decomposition).

This measure is also a variance decomposition and is computed similarly as for lmg but with weights attached to each single permutation:

$$pmvd(j) = \frac{1}{p!} p(r) \sum_r \Delta_j(r) \tag{5}$$

For more details about PMVD see Feldman [7] and also Grömping [2].

4.4 Relative Weights

Fabbris [8] has proposed a way to decompose the explained variance using the Singular Value Decomposition of the matrix X. Later Genizi [9] and Johnson [1] used this approach in a different way. This decomposition is a particular case of a more general approach consisting of using a set of mutually uncorrelated variable to decompose the explained variance. We will formalize the orthogonal decomposition in general and then present the Relative Weights computation.

Let $z_i, i = 1, \ldots p$ as set of p orthogonal standardized predictors:

Let us note $\lambda_{ji} = \text{cov}(z_j, X_i)$ and $\beta_i = \text{cov}(y, z_i)$

We compute the Orthogonal Decomposition RW with the zi's as follows

$$RW(j) = \sum_{i=1}^{i=p} \lambda_{ji}^2 \beta_i^2 \tag{6}$$

The Relative Weights generate a full variance decomposition because:

$$\sum_{j=1}^{j=p} RW(j) = \sum_{j=1}^{j=p} \sum_{i=1}^{i=p} \lambda_{ji}^2 \beta_i^2 = \sum_{i=1}^{i=p} \beta_i^2 \sum_{j=1}^{j=p} \lambda_{ji}^2 \tag{7}$$

$\lambda_{lm} = \text{cov}(z_l, x_m)$, and the z_i being a set of standardized orthogonal vectors and as the x_j are also standardized finally:

$$\sum_{i=1}^{i=p} RW(\text{i}) = V(y) = 1 \tag{8}$$

So the Relative Weights computed with any set of z_i enables the computation of a full decomposition of the R^2 using the RW_j. The decomposition proposed by Genizi and Johnson consists in computing the Relative Weights using a specific set of orthogonal predictors that minimize the sum of the squares between each X_j and z_j.

So in terms of variables minimizing: $\Psi = E[(z - X)'(\text{z} - X)]$

In the case of a specific dataset with n observations and p predictors this leads to consider a specific matrix Z of the zj is as follows:

Let X be the n x p matrix of standardized centered observations. Let $X = P\Delta Q'$ the singular value decomposition of X. The set of zi minimizing the abovementioned sum of squares is $Z = PQ'$.The orthogonal decomposition using this specific set of orthogonal vectors are the Relative Weights. We will use the notation RW(j) from now on for this specific decomposition and Vo(j) in case we use another set of zi's.

4.5 CAR scores

The CAR scores are the squared correlations between the response and the vectors Z as defined in 4.4.So: CAR(j)=λj²

This is a recent Variance Decomposition proposed by Zuber and Strimmer [10]. They use the term CAR standing for Correlation-Adjusted (marginal) coRelation.

We have limited the presentations of these methods to the strict minimum detail, but the documents in reference offer additional perspective on the Game Theory approach and axiomatic definitions of desirable properties in variance decomposition.

5 Results on Variance Decomposition

There are important difference between the usage of the Linear Model and the interpretation of Variance Decomposition values. In the case of lmg for instance, it is important not to use in a simplistic way the variance decomposition as if they were equivalent to the coefficients generated by Linear Regression Model. First because they are terms of variance that are actually homogeneous to squared values of the β's. If we consider an ideal case with mutually decorrelated predictors the lmg value would be distorted compared to the relative values generated by the Linear Model.
Also another way to write lmg in the case of two predictors is as follows:.

$\text{lmg}(1) = \frac{V(y) + (\beta_1^2 * \text{v}_1 - \beta_2^2 * \text{v}_2) * (1 - \rho_{12}^2)}{2}$ $\text{lmg}(2) = \frac{V(y) + (\beta_2^2 * \text{v}_2 - \beta_1^2 * \text{v}_1) * (1 - \rho_{12}^2)}{2}$ If we consider a case with two predictors a sufficiently high correlation and a third predictor uncorrelated with the first two the reallocation of importance between the two lmg values can lead ultimately to different rankings between the importance measures if we apply the beta squared versus lmg.

So all in all lmg produces distortion and can potentially change the ranking between the importance of predictors versus the results derived from the Linear Model. This is why we need to be careful not to consider them as a full alternative to standard

models. Conklin and Lipovetsky [11] have considered adjusting regression coefficients using Shapley Value as an intermediate step of calculation and computing coefficients in resolving a quadratic equation equalizing the Hoffman values and lmg for each predictor. Grömping and Landau [12] have criticized this approach.

Regarding similarities, Johnson and Lebreton [3] have observed the proximity between the results of relative weights and lmg and state :« *Despite being based on entirely different mathematical models, Johnson's epsilon and Budescu's dominance measures (Note : Budescu's dominance is one of the denomination of Shapley Value /lmg) provide nearly identical results when applied to the same data these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors. The convergence between these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors".*

We will first analyze and formulate results in the case of two predictors. Starting with two uncorrelated standardized variables E1 and E2, we will construct X1, X2, and y:

$$X_1 = \cos(\varphi)E_1 - \sin(\varphi)E_2 \qquad X_2 = \cos(\varphi)E_1 + \sin(\varphi)E_2$$
$$y = \cos(\psi)E_1 + \sin(\psi)E_2$$

From this we can actually compute:

$$\beta_1 = \frac{\sin(\varphi-\psi)}{\sin(2\varphi)} \qquad r_{y1} = \cos(\varphi+\psi)$$
$$\beta_2 = \frac{\sin(\varphi+\psi)}{\sin(2\varphi)} \qquad r_{y2} = \cos(\varphi-\psi)$$
$$\rho_{12} = \cos(2\varphi)$$

$$\text{last}(1) = \sin^2(\varphi-\psi); \text{first}(1) = \cos^2(\psi+\varphi)$$
$$\text{last}(2) = \sin^2(\psi+\varphi); \text{first}(2) = \cos^2(\psi-\varphi)$$

$$\text{SV}(1) = \frac{(1-\sin(2\varphi)\sin(2\psi))}{2}$$
$$\text{SV}(2) = \frac{(1+\sin(2\varphi)\sin(2\psi))}{2}$$

It is also possible to compute the result of orthogonal decomposition using any orthogonal base of the considered plane let us consider z1 and z2 such as:

$$z_1 = \cos(\omega)E_1 + \sin(\omega)E_2$$
$$z_2 = -\sin(\omega)E_1 + \cos(\omega)E_2$$

The results of an orthogonal decomposition process using the zi defined by the choice of a specific value of ω are Vo(1) and Vo(2) as computed below :

$$Vo(1) = \cos^2(\psi-\omega)\cos^2(\omega+\varphi) + \sin^2(\psi-\omega)\cos^2(\omega-\varphi)$$
$$Vo(2) = \cos^2(\psi-\omega)\sin^2(\omega+\varphi) + \sin^2(\psi-\omega)\sin^2(\omega-\varphi)$$

It is easy to demonstrate (w.l.o.g with φ≤π/2),that the specific z_i considered earlier to implement the Relative Weights of the variance decomposition proposed by Johnson and Genizi, corresponds to the case when ω=-π/4. Taking ω=-π/4 in the computations of Vo(1) and Vo(2) and simplifying we get:

$$\text{RW}(1) = \frac{(1-\sin(2\varphi)\sin(2\psi))}{2}; \text{RW}(2) = \frac{(1+\sin(2\varphi)\sin(2\psi))}{2}$$

We recognize here the formula for $SV(1)$ and $SV(2)$. So we have demonstrated through a trigonometric approach that in the case of two predictors the relative weights and the lmg (or Shapley Values) are identical (cf. also Thomas and al. [14]). This result is just the application of simple trigonometric equivalences and should not in our view lead to conclude that because the two methods converge this is in itself a justification of their validity. The demonstration proposed here enables easy visualization of the impact of the choice of orthogonalisation if we let ω vary. In the case of two predictors it also enables to demonstrate that the CAR scores may remain constant even when the correlation between predictors vary. We can also notice some other links between orthogonalisation procedures and lmg if we use some particular sets of orthogonal vectors in the

space generated by the Xj's. As Relative Weights is a particular case of decomposition by orthogonalisation , these links help understand the proximity between lmg which is an averaging of last values over submodels and relative weights , which is a decomposition by orthogonalisation.

Case 1: Let us consider y* as the projection of y on the space of the predictors and let us choose one given predictor and an orthogonal set of zi's with the condition that:

$$zj = \frac{y^*}{\| y^* \|}$$

We have

$$\forall i \neq j, y.zi = 0 \quad and \quad y^*.zj = 1$$

Let us now use the RW calculation formula:

$$\mathrm{Vo(j)} = \sum_{i=1}^{i=p} \lambda^2 \mathrm{ji} * \beta^2 \mathrm{i} \text{ as:}$$

$$\lambda ji = \mathrm{cov(zj}, Xi) \quad \beta i = \mathrm{cov(y, zi)} \quad \beta j \neq 0; \beta j = 1$$

we have :

$$\mathrm{Vo(j)} = \mathrm{cov}^2(\mathrm{y}, \mathrm{xj}) = \mathrm{first(j)}$$

This means that for any j there is always at least one choice of orthogonal decomposition that will allocate first (j) to that predictor.

Case 2: This time we will consider an orthogonal set so that:

$$zj = \frac{uj}{\| uj \|}$$

u_j being the residual of the regression of X_j on the other variables.

$$if \quad i \neq j \quad \lambda ji = \mathrm{cov(zj}, Xi) = 0, \quad and \quad \lambda^2 jj = \mathrm{cov}^2(\mathrm{z}\, j, Xj) = 1 - \mathrm{R}_j^2$$

$$As \quad \beta^2 j = \mathrm{cov}^2(\mathrm{y}, \mathrm{zj}), \quad \mathrm{Vo(j)} = last(j)$$

These examples show that orthogonalisation methods do enable with specific sets of orthogonal vectors to allocate either first(j) or last(j) for one given predictor. As Johnson is a particular case of orthogonalisation and lmg (j) is an average of last (j) across submodels, it confirms why there can be a proximity between variance decomposition via orthogonalisation and lmg in the case of more than 2 predictors. Both lmg and Relative Weights are computer intensive methods. We introduce an alternative variance decomposition method that is much more computationally efficient and offers very similar results to lmg and relative weights. The method will allocate to each predictor j a share of variance that is a weighted average between first(j) and last (j) , hence the name weifila for weighted first last.

Here are the computation steps: let L and F be the sum of first and last for all predictors:

$$L = \sum_j last(j) \qquad F = \sum_j \mathrm{first}(j)$$

We will consider two cases in the usual situation where $L \neq F$

$$If \quad L < R^2 < F \quad \mathrm{W}(j) = last(j)\left(\frac{F - R^2}{F - L}\right) + first(j)\left(\frac{R^2 - L}{F - L}\right) \tag{9}$$

$$If \quad F < R^2 < L \quad \mathrm{W}(j) = last(j)\left(\frac{R^2 - F}{L - F}\right) + first(j)\left(\frac{L - R^2}{L - F}\right) \tag{10}$$

The case where R² would be outside of the interval between F and L is not encountered in practice. By construction in both cases above: $\sum w(j) = R^2$

We will note also that in the case of two predictors the weifila values equate to the lmg and relative weights values as shown below:

$$F = first(1) + first(2) = 1 + \cos(2\psi)\cos(2\varphi)$$

$$L = last(1) + last(2) = 1 - \cos(2\psi)\cos(2\varphi) \quad F - L = 2\cos(2\varphi)\cos(2\psi) \qquad w(j) = \frac{first(\mathrm{j}) + \mathrm{last(j)}}{2}$$

So we recognize the formula above for lmg and this confirms that: $w(j) = lmg(j) = SV(j) = RW(j)$ (11)

Weifila is a way to select an intermediate point w(j) inside the interval between first(j) and last(j) for each j. We have compared these 3 measures on two different datasets. The results are presented below:

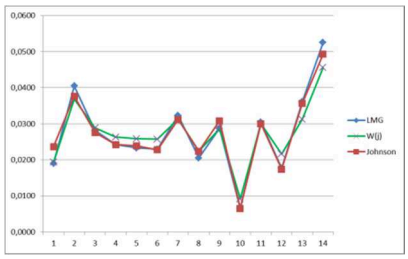


Fig 1: 1499 obsevations.14 predictors

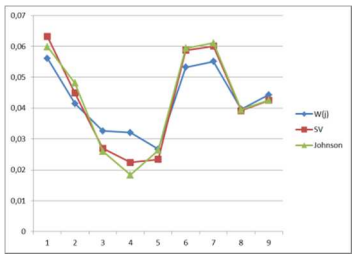


Fig 2: 499 observations 9 predictors

The weighted first last average "weifila" is much simpler to compute and delivers very similar results at least in the typical size of datasets and number of drivers used in practical applications for marketing and social research.

6 Conclusions

Among several methods to allocate variance among predictors, the proximity between lmg and Relative Weights has been noted (Johnson and Lebreton [3]), and seen as a justification of their validity. This proximity is actually a complete equality in the case of a model with two predictors and results from simple geometric properties. Also there exists variance decomposition via orthogonalisation that allocate exactly last(j) or first(j) for any of the predictors. So this proximity should not be seen in itself as a justification of validity.

The new method of variance decomposition proposed in this paper via a weighted average between first(j) and last(j) for each predictor provides very consistent results with lmg and Relative Weights but is simpler and less computer intensive. This method has been successfully tested with datasets typical of situations encountered in marketing research applications.

As underlined by Johnson [1] and Grömping [2], variance decomposition should not be seen as a substitute for linear regression models or path analytical models and models based on theory driven explanations can be more relevant than using directly variance decomposition. However when a model based on theory is not available variance decomposition can help identify important variables. Lastly the usage of modern machine learning techniques can also be considered.

References

1. J. W.Johnson. A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression. Multivariate Behavioral Research,35(1) 1-19. 2000.
2. U.Grömping. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. The American Statistician, Vol 61 , No2 p139 2007.
3. J.W Johnson and J.M. Lebreton. History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods* **7**, 238 - 257. . 2004.
4. http://prof.beuth-hochschule.de/groemping/relaimpo/
5. V.Zuber, and K.Strimmer. Variable importance and model selection by decorrelation. Preprint. http://arxiv.org/abs/1007.5516 (2010).
6. B. Tabachnick L. Fidell , Using Multivariate Statistics , Fifth Edition , 5.6.1.1 Pearson 2006.
7. B Feldman . "Relative Importance and Value." Manuscript version 1.1, 2005 http://www.prismanalytics.com/docs/RelativeImportance.pdf
8. L. Fabbris. Measures of regressor importance in multiple regression: an additional suggestion. *Qual Quant* 1980, 4:787–792
9. A. Genizi. Decomposition of R^2 in Multiple Regression with correlated regressors. Statiscica Sinica 4 , 407-420. 1993.
10. V. Zuber, K Strimmer. High-Dimensional Regression and Variable Selection Using CAR Scores. Statistical Applications in Genetics and Molecular Biology **10**: 34. 2011.
11. S.Lipovetsky, M. Conklin. Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry* **17**, 319-330.2001
12. U Grömping , S.Landau.Do not adjust coefficients in Shapley Value Regression *Applied Stochastic Models in Business and Industry* **17**, 319-330.2009
13. U Grömping. Variable importance in regression models *WIRE's Comput Stat 2015,7:137-152. Doi:10.1002/wics.1346*
14. R D Thomas , Bruno D Zumbo, Ernest Kwan, Linda Schweitzer. On Johnson's (2000) Relative Weights Method for Assessing Variable Importance A Reanalysis : *Multivariate Behavioral Research July 18 2014*

Bibliographie

Achen, C. (1982) Interpreting and using regression. *Series Quantitative Applications in the Social Sciences*. Sage University Paper. Volume 29.

Ardilly, P. (2006) *Les techniques de Sondage*. Editions Technip.

Azen, R., Budescu, D.V. (1993). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods* **8**, 129-148.

Azen, R. (2003). Dominance Analysis SAS Macros. URL: www.uwm.edu/~azen/damacro.html.

Bollen, K.A. (1989). *Structural equations with latent variables.* Wiley, New York.

Braun, M. T., & Oswald, F. L. (2011). Exploratory regression analysis : A tool for selecting models and determining predictor importance. *Behavior Research Methods*, *43*, 331–339.

Breiman, L. (2001), Random Forests, *Machine Learning, 45:5-32.*

Brown, L., Tsamardinos, I., Aliferis, C.F.(2005) A Comparison of Novel and State-of-the-Art Polynomial Bayesian Network Learning Algorithms. *American Association for Artificial Intelligence. http://www.cs.mtu.edu/~lebrown/papers/brown.etal.aaai.2005.pdf*

Budescu, D.V. (1993). Dominance Analysis: A new approach to the problem of relative importance in multiple regression. *Psychological Bulletin* **114**, 542-551.

Budescu, D.V. and Azen, R. (2004). Beyond Global Measures of Relative Importance: Some Insights from Dominance Analysis. *Organizational Research Methods* **7**, 341 - 350.

Bühlmann,P.(2013) Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research*. Springer.77:357-370 DOI 10:1007/s00186-012-0404-7

Bühlmann,P.,Maathuis,M.H.,Colombo,D.,Mächler,M.,Kalish;M. (2012) Causal Inference Using Graphical Models with the R package pcalg. *Journal of Statistical Software. April 2012, Volume 47, Issue 11.*

Chieng, J. (1998). Belief Network Project Constructor. https://webdocs.cs.ualberta.ca/~jcheng/bnpc.htm

Cohen, J.,Cohen,P., West,S.G., Aiken,L.S.,(1975et 2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associate Publishers , Mahwah New Jersey ISBN-13: 978-0805822236.

Colias, J.(2007). New Statistical Tools for Key Driver Analysis. Publication by Decision Analyst. www.decisionanalyst.com

Conklin, M, Lipovetsky, S. (2013). The Shapley Value in Marketing Research : 15 years and counting *Proceedings of the Sawtooth Conference October 2013*.http://www.sawtoothsoftware.com/support/technical-papers/conference-proceedings/100-support/proceedings/1426-proceedings2013

Cosnefroy, O., Sabatier, C. (2011).Estimation de l'importance relative des prédicteurs dans un modèle de régression multiple. Intérêt et limites des méthodes récentes. L'Année psychologique, 111, pp 253-289 doi:10.4074/S000350331100202

Courville.T, Thompson, B.(2001) Use of Structure Coefficients in Published Multiple Regression Articles: β is not Enough. *Educational and Psychological Measurement April 2001 61: 229-248,*

Darlington., R.B. (1968) Multiple Regression in Psychological Research and Practice. *Psychological Bulletin,69. 161_182.*

Denis. J-B, Scutari, M. (2014) *Réseaux Bayesiens avec R.* Editions EDP Sciences. Paris.

Egner, M. Hart, R. (2012) Modifying Bayesian Networks for Key Drivers Analysis: An overview of practical improvements, Sawtooth Conference 2012 (Proceedings, pages 361-373).

Eidelman, A. (2012). La valeur de Shapley : Comment individualiser le résultat d'un groupe http://www.insee.fr/fr/themes/document.asp?reg_id=0&ref_id=F1202

Darlington, R.B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin 69, 161-182.

Engelhart.M.,(1936) The technique of path coefficients. *Psychometrika.*1(4):287-293. DOI: 10.1007/BF02287881

Fabbris L. (1980) Measures of regressor importance in multiple regression: an additional suggestion. *Qual Quant* 1980*, 4:787–792,*

Feldman, B. (1999). The proportional value of a cooperative game. *Manuscript for a contributed paper at the Econometric Society World Congress 2000*. Downloadable at http://fmwww.bc.edu/RePEc/es2000/1140.pdf.

Feldman, B. (2005). *A Dual Model of Cooperative Value*. Manuscript, downloadable from http://ssrn.com/abstract=31728Feldman, B. Relative Importance and Value. Manuscript (latest version), downloadable at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2255827.

Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica* 3, 407-420.

Genuer, R., Poggi, J-M., Tuleau, C. Random Forests: some methodological insights. [Research Report] RR-6729, 2008.<inria-00340725>. (**arXiv:0811.3619v1**)

Genuer, R., Poggi, J-M., Tuleau-Malot, C. Variable selection using Random Forests. *Pattern Recognition Letters,* Elsevier, 2010, 31 (14), pp.2225-2236.

Gibson., W.A. (1962) Orthogonal Predictors : a possible resolution of the Hoffman-Ward controversy. *Psychological Reports:* Volume 11, Issue. pp. 32-34.

Giannelloni, J-L., Vernette, E .(2012) *Etudes de marché*. Editions Vuibert.

Grapentine, T. (2012) *Applying Scientific Reasoning to the Field of Marketing: Make Better Decisions* (Marketing Strategy Collection). Business Expert Press.

Green, P.E., Tull, D.S. (1975) *Research for Marketing Decisions*; 3rd edition; Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

Gregorutti, B., Michel, B., Saint-Pierre, P. (2013). Correlation and variable importance in random forests. axXiv :1310.5726.

Gregorutti, B., Michel, B., Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. . axXiv :1411.4170v2

Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software* 17, Issue 1.

Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician* 61: 139-147.

Grömping, U. (2007). Response to comment by Scott Menard, re: Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. In: Letters to the Editor, *The American Statistician* 61: 280-284.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression Versus Random Forest. *The American Statistician* 63: 308-319.

Grömping, U., Landau, S. (2009). Do not adjust coefficients in Shapley value regression. *Applied Stochastic Models in Business and Industry*. Wiley. DOI: 10.1002/asmb.773.

Grömping, U. (2015). Variable Importance in Regression *Models WIREs Comput Stat. (*2015). 7:137–152. doi: 10.1002/wics.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Wiley.

Hoffman, P.J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin* **57**, 116-131.

Hüttner, F. and Sunder, M. (2011). Decomposing R2 with the Owen Value.*Working Paper No 100.ISSN 1437-9384 Faculty of Economics and Management Science. Université de Leipzig.*

Hüttner, F. and Sunder, M. (2012). Stata module for decomposing goodness of fit according to Owen and Shapley. Stata:*htttp://www.stata.com/meeting/uk12/abstracts/.*

Huettner, F. and Sunder, M. (2012). Axiomatic arguments for decomposing goodness to fit according to Shapley and Owen Values. *Electronic Journal of Statistics.* Vol. 6.1239–1250 ISSN: 1935-7524 DOI: 10.1214/12-EJS710

IBM SPSS (2010). How to get more value from your survey data. *Site ibm.com.* http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=YTW03033GBEN aussi disponible sur le site *kdnuggets* : http://www.kdnuggets.com/2012/08/ibm-how-to-get-more-value-from-your-survey-data.html

Iooss, B., Lemaître, P. (2014). A review on global sensitivity analysis methods. *arXiv:1404.2405v1 [math.ST] 9 Apr 2014*.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics. Vol 1.519-537.*

Johnson, J.W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research* 35, 1-19.

Johnson, J.W. and Lebreton, J.M. (2004). History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods* **7**, 238 - 257.

Jöreskog, K.G., Wold, H. (1982) *The ML and PLS techniques for modelling with latent variables: historical and competitive aspect*, en Jöreskog, K.G. et Wold, H. (Editors), Systems under indirect observation, Part 1, pages 263-270, North-Holland, Amsterdam.

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions.* Sage Publications Inc. Thousands Oaks, California.

Kenett, R. Salini, S.(2012). *Modern Analysis of Customer Surveys with applications using R.* Wiley. Londres.

Kruskal, W., Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician* **43**: 2-6.

Leray, P. (2006). Réseaux Bayésiens: Apprentissage et diagnostic de systemes complexes. Modeling and Simulation. Université de Rouen, 2006. *tel-00485862. https://tel.archives-ouvertes.fr/tel-00485862/document*

Lebreton, J.M. Ployhart, R.E. and Ladd, R.T. (2004). A Monte Carlo Comparison of Relative Importance Methodologies. *Organizational Research Methods* **7**, 258 - 282.

Lindeman, R. H., Merenda, P.F., Gold, R.Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman, Glenview IL.

Lipovetsky, S. and Conklin, M. (2001). Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry* **17**, 319-330.

McLauchlan, W.G. (1992). Regression-based satisfaction analyses: proceed with caution. *Quirk's Marketing Research Review, October, 1992, p. 10-13*

Marley, Anthony AJ; Louviere, Jordan J. (2005-01-01). Some probablistic models of best, worst, and bewt-worst choices. *Journal of Mathematical Psychology* **49** (6): 464–480.

Martilla, J.A., James, J.C. (1977) Importance-Performance Analysis. *Journal of Marketing. Vol. 41, No. 1 (Jan., 1977), pp. 77-79. DOI: 10.2307/1250495.*

Menard, S. (2007). Letter to the Editor, *The American Statistician, August 2007, Vol. 61, No. 3 , p 280.*

Naïm, P., Wuillemin, P.H., Leray, P., Pourret, O.,Becker, A. (2004). *Réseaux Bayésiens.*Editions Eyrolles.Paris.

Nathans,L., Oswald,F.L., Nimon, K. C. (2012). Interpreting Multiple Linear Regression: a Guidebook of Variable Importance. *Practical Assessment, Research and Evaluation.Volume 17, Number 9, April 2012.ISSN 1531-7714.*

Nimon K.F., Oswald F.L (2013). Understanding the resultsof multiple linear regression: beyond standardized regression coefficients. *Organ Res Methods* 2 16: 650–674

Owen A.B , (2014). Sobol's indices and Shapley value. *Society of Industrial and Applied Mathematics (SIAM) 2014.*

Pearl, J. Causality: *Models Reasoning and Inference*. Cambridge University Press (2009).

Pokryshevskaya,E.,Antipov,E. (2013) The comparison of methods used to measure the importance of service attributes. *National Research University Higher School of Economics, Department of Economics Saint-Petersburg*, Russia. *http://ssrn.com/abstract=2215680*

Pratt, J.W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In: Pukkila, T. and Puntanen, S. (Eds.): *Proceedings of second Tampere conference in statistics*, University of Tampere, Finland, 245-260.

Robinson,R. (1977). Counting unlabeled acyclic diagraphs. In C. Little, editor, *Combinatorial Mathematics V,* volume 622 of *Lecture Notes in Mathematics*, page 28-43,Berlin,1997. Springer.

Saltelli, A.,Tarantola, S.,Campolongo,F.,Ratto,M. (2005) *Sensitivity Analysis in Practice*. Wiley.

Saltelli, A.,Tarantola,S.,Campolongo,F.,Ratto,M.Andres,T.,Cariboni,J.,Gatelli,D.,Saisana,M.(2008).*Global* S*ensitivity*. Wiley .

Shear, B.,Olvera,O., Zumbo,B.D. (2012). Relative variable importance for multiple regression in suppressor situations: A Monte Carlo study. *Paper presented at the American Educational Research Association annual meeting, Vancouver, British Columbia, Canada*

Schafer, W.D. (1991). Reporting hierarchical regression results. *Measurement and Evaluation in Counseling and Development 24, 98-100*.

Shapley, L. (1953). A value for n-person games. Reprinted in: Roth, A. (1988, ed.): *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, Cambridge.

Schmitt, N. (1996). Uses and Abuses of Coefficient alpha. *Psychological Assessment*. *The American Psychological Association, Inc. 1996*, Vol. 8, No. 4, 350-353

Song E., Nelson B.L, Staum J. (2014). Shapley Effects for Global Sensitivity Analysis. *Nothwestern University September 3 , 2014*.

Spirtes, P.,Glymour,C.,Scheines,R. Causation, Prediction, and Search. *Department of Philosophy Carnegie Mellon University (SGS, 1993)*

Strobl, C., Boulesteix, A-L., Zeileis, A., Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, [p2].

Strobl, C., Boulesteix,A-L.,Kneib,T.,Augustin,T.,Zeileis , A. Conditional variable importance for random forests. *BMC Bioinformatics 2008, **9**:307 doi:10.1186/1471-2105-9-307*

Strimmer, K. Zuber, V. High-Dimensional Regression and Variable Selection Using CAR Scores. Statistical Applications in Genetics and Molecular Biology **10**: 34. 2011.

Tabachnick, B.,Fidell.L.,*Using Multivariate Statistics*, Pearson (2001).ISBN-13: 978-0205849574

Thomas, D.R., Hughes, E. and Zumbo, B.D. (1998). On variable importance in linear regression. *Social Indicators Research* **45**, 253-275.

Thomas, D.R., Zhu, P.C. and Decady, Y.J. (2007). Point estimates and confidence intervals for variable importance in multiple linear regression. *J. Educational and Behavioral Statistics* **32**, 61-91.

Thomas D.R, Zumbo, B.D. Kwan, E.,Schweitzer,L.(2014) On Johnson's (2000) Relative Weights Method for Assessing Variable Importance: A Reanalysis. *Multivariate Behaviora l Research*, 49:329–338, 2014

Ticehurst, J.L.,Curtis,A.,Merritt,W.S. (2010) Using Bayesian Network s to complem ent conventional analyses to explore landholder management of native vegetation. *Environment Modeling and Software*. Elsevier. doi:10.1016/j.envsoft.2010.03.032

Tonidandel, S., Lebreton, J. (2011). Relative Importance Analysis: A Useful Supplement to Regression Analysis. *Journal of Business and Psychology*. 26:1–9 DOI 10.1007/s10869-010-9204-3

Tufféry, S. (2015). *Modélisation predictive et Apprentissage statistique*. Editions TECHNIP. Paris.

Verma, T. Pearl, J. (1991) Equivalence and synthesis of causal models. In M. Henrion, R. Shachter,L. Kanal, and J. Lemmer, editors, *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence,* pages 220-227, San Francisco.

Vuong (1989) Likelihood Ratio Tests for Model Selection and Non Nested Hypotheses. Econometrica Volume 57 Issue 2 Mar., 1989, 307-333

Wallard, H. (2015) Using Explained Variance Allocation to analyse Importance of Predictors. *The 16th Conference of the Applied Stochastic Models and Data Analysis International Society. (Piraeus,Greece). http://www.asmda.es/images/1_W-Z_ASMDA2015_Proceedings.pdf.*

Wold, H. (1975) *Soft modeling by latent variables : The Non-Linear Iterative Partial Least Squares (NIPALS) approach,* in Gani, J. (Editor), Perspectives in probability and statistics, pages 117-142, Londres.

Wold, H. (1985) *Partial Least Squares,* Kotz S. et Johnson, N.L. (Editors), Encyclopedia of Statistical Sciences, Vol 6, pages 581-591, John Wiley & Sons , New York .

Zheng, Z. Pavlou,P. (2009)Towards a Causal Interpretation from Observational Data : A New Bayesian Networks Method for Structural Models with Latent Variables. *Information Systems Research ISSN* 1047-7047/EISSN 1526-5536

Zuber, V. (2010). High Dimensional Variable Selection by decorrelation: introducing the CAR scores. *https://www.wias-berlin.de/workshops/validation2010/workshop_slides/talk_zuber.pdf*

Résumé.

La colinéarité rend difficile l'utilisation de la régression linéaire pour estimer l'importance des variables dans les études de marché. D'autres approches ont donc été utilisées.

Concernant la décomposition de la variance expliquée, une démonstration de l'égalité entre les méthodes *lmg-Shapley* et celle de Johnson avec deux prédicteurs est proposée. Il a aussi été montré que la méthode de Fabbris est différente des méthodes de Genizi et Johnson et que les *CAR scores* de deux prédicteurs ne s'égalisent pas lorsque leur corrélation tend vers 1.Une méthode nouvelle, *weifila* (*weighted first last*) a été définie et publiée en 2015.

L'estimation de l'importance des variables avec les forêts aléatoires a également été analysée et les résultats montrent une bonne prise en compte des non-linéarités.

Avec les réseaux bayésiens, la multiplicité des solutions et le recours à des restrictions et choix d'expert militent pour utilisation prudente même si les outils disponibles permettent une aide dans le choix des modèles.

Le recours à *weifila* ou aux forêts aléatoires est recommandé plutôt que *lmg-Shapley* sans négliger les approches structurelles et les modèles conceptuels.

Mots clés :

régression, décomposition de la variance, importance, valeur de Shapley, forêts aléatoires, réseaux bayésiens.

Abstract

Linear regression is used in Market Research but faces difficulties due to multicollinearity. Other methods have been considered.

A demonstration of the equality between *lmg-Shapley* and and Johnson methods for Variance Decomposition has been proposed. Also this research has shown that the decomposition proposed by Fabbris is not identical to those proposed by Genizi and Johnson, and that the *CAR scores* of two predictors do not equalize when their correlation tends towards 1. A new method, *weifila (weighted first last)* has been proposed and published in 2015.

Also we have shown that permutation importance using Random Forest enables to take into account non linear relationships and deserves broader usage in Marketing Research.

Regarding Bayesian Networks, there are multiple solutions available and expert driven restrictions and decisions support the recommendation to be careful in their usage and presentation, even if they allow to explore possible structures and make simulations.

In the end, *weifila* or random forests are recommended instead of *lmg-Shapley* knowing that the benefit of structural and conceptual models should not be underestimated.

Keywords :

Linear regression, Variable Importance, Shapley Value, Random Forests, Bayesian Networks.